ICINCO 2009

6TH INTERNATIONAL CONFERENCE ON INFORMATICS IN CONTROL, AUTOMATION AND ROBOTICS

Proceedings

Volume 3 - Signal Processing, Systems Modeling and Control

MILAN - ITALY · JULY 2 - 5, 2009

ORGANIZED BY
INSTICC

IN COOPERATION WITH
aaai

CO-SPONSORED BY
IFAC

INSTICC PRESS

ICINCO 2009

Proceedings of the
6th International Conference on
Informatics in Control, Automation and Robotics

Volume 3
Signal Processing, Systems Modeling and Control

Milan, Italy

July 2 - 5, 2009

Organized by
INSTICC – Institute for Systems and Technologies of Information, Control and Communication

Co-Sponsored by
IFAC – International Federation of Automatic Control

In Cooperation with
AAAI – Association for the Advancement of Artificial Intelligence

Edited by Joaquim Filipe, Juan Andrade Cetto and Jean-Louis Ferrier

Printed in Portugal

ISBN: 978-989-674-001-6

Depósito Legal: 294961/09

http://www.icinco.org

secretariat@icinco.org

Brief Contents

INVITED SPEAKERS

Daniel S. Yeung

University of Technology

China

Maria Pia Fanti

Polytechnic of Bari

Italy

Janan Zaytoon

University of Reims Champagne Ardennes

France

Alessandro Giua

University of Cagliari

Italy

Peter S. Sapaty

Institute of Mathematical Machines and Systems, National Academy of Sciences

Ukraine

ORGANIZING AND STEERING COMMITTEES

CONFERENCE CHAIR

Joaquim Filipe, Polytechnic Institute of Setúbal / INSTICC, Portugal

PROGRAM CHAIR

Juan Andrade Cetto, Institut de Robòtica i Informàtica Industrial CSIC-UPC, Spain

Jean-Louis Ferrier, University of Angers, France

PROCEEDINGS PRODUCTION

Sérgio Brissos, INSTICC, Portugal

Helder Coelhas, INSTICC, Portugal

Vera Coelho, INSTICC, Portugal

Andreia Costa, INSTICC, Portugal

Bruno Encarnação, INSTICC, Portugal

Bárbara Lima, INSTICC, Portugal

Raquel Martins, INSTICC, Portugal

Carla Mota, INSTICC, Portugal

Vitor Pedrosa, INSTICC, Portugal

José Varela, INSTICC, Portugal

CD-ROM PRODUCTION

Elton Mendes, INSTICC, Portugal

Pedro Varela, INSTICC, Portugal

GRAPHICS PRODUCTION AND WEBDESIGNER

Marina Carvalho, INSTICC, Portugal

SECRETARIAT AND WEBMASTER

Marina Carvalho, INSTICC, Portugal

PROGRAM COMMITTEE

Arvin Agah, The University of Kansas, U.S.A.

Alessandro Chiuso, Universita di Padova, Italy

Hyo-Sung Ahn, Gwangju Institute of Science and Technology (GIST), Korea, Republic of

Frank Allgower, University of Stuttgart, Germany

Francesco Amigoni, Politecnico di Milano, Italy

Plamen Angelov, Lancaster University, U.K.

Peter Arato, Budapest University of Technology and Economics, Hungary

Helder Araújo, University of Coimbra, Portugal

Marco Antonio Arteaga, Universidad Nacional Autonoma de Mexico, Mexico

Vijanth Sagayan Asirvadam, Universiti Technologi PETRONAS, Malaysia

T. Asokan, Indian Institute of Technology Madras, India

Robert Babuska, Delft University of Technology, The Netherlands

Ruth Bars, Budapest University of Technology and Economics, Hungary

Adil Baykasoglu, University of Gaziantep, Turkey

Maren Bennewitz, University of Freiburg, Germany

Karsten Berns, University of Kaiserslautern, Germany

Arijit Bhattacharya, Dublin City University, Ireland

Sergio Bittanti, Politecnico di Milano, Italy

Stjepan Bogdan, University of Zagreb, Faculty of EE&C, Croatia

Jean-louis Boimond, ISTIA - LISA, France

Djamel Bouchaffra, Oakland University, U.S.A.

Bernard Brogliato, INRIA, France

Edmund Burke, University of Nottingham, U.K.

Clifford Burrows, Innovative Manufacturing Research Centre, U.K.

Dídac Busquets, Universitat de Girona, Spain

Giuseppe Carbone, LARM - Laboratorio di Robotica e Meccatronica, Italy

J. L. Martins de Carvalho, Instituto de Sistemas e Robótica - Porto, Portugal

Alessandro Casavola, University of Calabria, Italy

Riccardo Cassinis, University of Brescia, Italy

Chien Chern Cheah, Nanyang Technological University, Singapore

Tongwen Chen, University of Alberta, Canada

Wen-Hua Chen, Loughborough University, U.K.

Graziano Chesi, University of Hong Kong, China

Carlos Coello Coello, Cinvestav-IPN, Mexico

Yechiel Crispin, Embry-riddle Aeronautical University, U.S.A.

Michael A. Demetriou, Worcester Polytechnic Institute, U.S.A.

Guilherme DeSouza, University of Missouri, U.S.A.

Jorge Dias, Institute of Systems and Robotics, Portugal

Gamini Dissanayake, University of Technology, Sydney, Australia

Denis Dochain, Université Catholique de Louvain, Belgium

Tony Dodd, The University of Sheffield, U.K.

Alexandre Dolgui, Ecole des Mines de Saint Etienne, France

Prabu Dorairaj, Wipro Technologies, India

Marco Dorigo, Université Libre de Bruxelles, Belgium

Venky Dubey, Bournemouth University, U.K.

Petr Ekel, Pontifical Catholic University of Minas Gerais, Brazil

Andries Engelbrecht, University of Pretoria, South Africa

Sebastian Engell, Univeristy of Dortmund, Germany

PROGRAM COMMITTEE (CONT.)

Simon G. Fabri, University of Malta, Malta

Sergej Fatikow, University of Oldenburg, Germany

Jean-marc Faure, Ecole Normale Superieure de Cachan, France

Paolo Fiorini, Università degli Studi di Verona, Italy

Juan J. Flores, University of Michoacan, Mexico

Georg Frey, German Researc Center for Artificial Intelligence - DFKI, Germany

Manel Frigola, Technical University of Catalonia (UPC), Spain

John Qiang Gan, University of Essex, U.K.

Nicholas Gans, National Reasearch Council and Air Force Research Laboratory, U.S.A.

Leonardo Garrido, Monterrey Institute of Technology, Mexico

Andrea Garulli, Universita' di Siena, Italy

Lazea Gheorghe, Technical University of Cluj-Napoca, Romania

Paulo Gil, Universidade Nova de Lisboa, Portugal

Alessandro Giua, University of Cagliari, Italy

Luis Gomes, Universidade Nova de Lisboa, Portugal

Dongbing Gu, University of Essex, U.K.

Guoxiang Gu, Louisiana State University, U.S.A.

Jason Gu, Dalhousie University, Canada

Wail Gueaieb, University of Ottawa, Canada

José J. Guerrero, Universidad de Zaragoza, Spain

Thomas Gustafsson, Luleå University of Technology, Sweden

Maki K. Habib, Saga University, Japan

Hani Hagras, University of Essex, U.K.

Wolfgang Halang, Fernuniversitaet, Germany

Riad Hammoud, Delphi Corporation, U.S.A.

Uwe D. Hanebeck, Institute for Anthropomatics, Germany

Robert Harrison, The University of Sheffield, U.K.

Dominik Henrich, University of Bayreuth, Germany

Francisco Herrera, University of Granada, Spain

Victor Hinostroza, University of Ciudad Juarez, Mexico

Wladyslaw Homenda, Warsaw University of Technology, Poland

Guoqiang Hu, Kansas State University, U.S.A.

Marc Van Hulle, K. U. Leuven, Belgium

Fumiya Iida, Robot Locomotion Group, U.S.A.

Atsushi Imiya, IMIT Chiba University, Japan

Hisao Ishibuchi, Osaka Prefecture University, Japan

Thira Jearsiripongkul, Thammasat University, Thailand

Dimitrios Karras, Chalkis Institute of Technology, Greece

Dusko Katic, Mihailo Pupin Institute, Serbia

Graham Kendall, University of Nottingham, U.K.

Tamas Keviczky, Delft University of Technology, The Netherlands

Jonghwa Kim, University of Augsburg, Germany

Won-jong Kim, Texas A&M University, U.S.A.

Waree Kongprawechnon, Thammasat University, Thailand

Israel Koren, University of Massachusetts, U.S.A.

George L. Kovács, Hungarian Academy of Sciences, Hungary

H. K. Lam, King's College London, U.K.

Kemal Leblebicio, Middle East Technical University, Turkey

Graham Leedham, University of New England, Australia

Kauko Leiviskä, University of Oulu, Finland

Kang Li, Queen's University Belfast, U.K.

Tsai-Yen Li, National Chengchi University, Taiwan

PROGRAM COMMITTEE (CONT.)

Yangmin Li, University of Macau, China

Huei-Yung Lin, National Chung Cheng University, Taiwan

Zongli Lin, University of Virginia, U.S.A.

Jing-Sin Liu, Institute of Information Science, Academis Sinica, Taiwan

Jose Tenreiro Machado, Institute of Engineering of Porto, Portugal

Frederic Maire, Queensland University of Technology, Australia

Om Malik, University of Calgary, Canada

Jacek Mandziuk, Warsaw University of Technology, Poland

Hervé Marchand, INRIA, France

Gerard Mckee, The University of Reading, U.K.

Seán McLoone, National University of Ireland (NUI) Maynooth, Ireland

Carlo Menon, Simon Fraser University, Canada

Sanya Mitaim, Thammasat University, Thailand

Pieter Mosterman, The MathWorks, U.S.A.

Rafael Muñoz-salinas, University of Cordoba, Spain

Kenneth Muske, Villanova University, U.S.A.

Andreas Nearchou, University of Patras, Greece

Luciana P. Nedel, Universidade Federal do Rio Grande do Sul (UFRGS), Brazil

Sergiu Nedevschi, Technical University of Cluj-Napoca, Romania

Anton Nijholt, University of Twente,
The Netherlands

Hendrik Nijmeijer, Eindhoven University of Technology, The Netherlands

Juan A. Nolazco-Flores, ITESM, Campus Monterrey, Mexico

Urbano Nunes, University of Coimbra, Portugal

José Valente de Oliveira, Universidade do Algarve, Portugal

Romeo Ortega, LSS/CNRS/Supélec, France

Manuel Ortigueira, Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa, Portugal

Selahattin Ozcelik, Texas A&M University-Kingsville, U.S.A.

Christos Panayiotou, University of Cyprus, Cyprus

Stefano Panzieri, Università degli Studi "Roma Tre", Italy

Igor Paromtchik, INRIA, France

D. T. Pham, Cardiff University, U.K.

Marie-Noëlle Pons, CNRS, France

Raul Marin Prades, Jaume I University, Spain

Jerzy Respondek, Silesian University of Technology, Poland

A. Fernando Ribeiro, Universidade do Minho, Portugal

Robert Richardson, University of Leeds, U.K.

Rodney Roberts, Florida State University, U.S.A.

Juha Röning, University of Oulu, Finland

António Ruano, CSI, Portugal

Fariba Sadri, Imperial College London, U.K.

Carlos Sagüés, University of Zaragoza, Spain

Mehmet Sahinkaya, University of Bath, U.K.

Antonio Sala, Universidad Politecnica de Valencia, Spain

Abdel-badeeh Salem, Ain Shams University, Egypt

Mitsuji Sampei, Tokyo Institute of Technology, Japan

Medha Sarkar, Middle Tennessee State University, U.S.A.

Jurek Sasiadek, Carleton University, Canada

Daniel Sbarbaro, Universidad de Concepcion, Chile

Carla Seatzu, University of Cagliari, Italy

João Sequeira, Instituto Superior Técnico / Institute for Systems and Robotics, Portugal

Program Committee (cont.)

Michael Short, University of Leicester, U.K.

Silvio Simani, University of Ferrara, Italy

Dan Simon, Cleveland State University, U.S.A.

Adam Slowik, Koszalin University of Technology, Poland

Michael Small, Hong Kong Polytechnic University, Hong Kong

Burkhard Stadlmann, University of Applied Sciences Wels, Austria

Tarasiewicz Stanislaw, Université Laval, Canada

Karl Stol, University of Auckland, New Zealand

Olaf Stursberg, Technische Universitaet Muenchen, Germany

Chun-Yi Su, Concordia University, Canada

Cornel Sultan, Virginia Tech, U.S.A.

Ryszard Tadeusiewicz, AGH University of Science and Technology, Poland

Choon Yik Tang, University of Oklahoma, U.S.A.

Daniel Thalmann, VR Lab EPFL, Switzerland

N. G. Tsagarakis, Istituto Italiano di Tecnologia, Italy

Antonios Tsourdos, Cranfield University (Cranfield Defence and Security), U.K.

Nikos Tsourveloudis, Technical University of Crete, Greece

Anthony Tzes, University of Patras, Greece

Dariusz Ucinski, University of Zielona Gora, Poland

Nicolas Kemper Valverde, Universidad Nacional Autónoma de México, Mexico

Eloisa Vargiu, University of Cagliari, Italy

Laurent Vercouter, Ecole Nationale Superieure des Mines de Saint-Etienne, France

Bernardo Wagner, University of Hannover, Germany

Axel Walthelm, sepp.med GmbH, Germany

Dianhui Wang, La Trobe University, Australia

Qing-Guo Wang, National University of Singapore, Singapore

Zidong Wang, Brunel University, U.K.

James Whidborne, Cranfield University, U.K.

Dirk Wollherr, Technische Universität München, Germany

Marek Zaremba, Université du Québec (UQO), Canada

Janan Zaytoon, University of Reims Champagne Ardennes, France

Qin Zhang, University of Illinois at Urbana-Champaign, U.S.A.

Auxiliary Reviewers

Andrea Baccara, Università degli Studi di Cagliari, Italy

Rui Cortesao, University of Coimbra, Portugal

Matteo de Felice, ENEA, Italy

Andrea Gasparri, University of Roma Tre, Italy

Zhi Han, The MathWorks, U.S.A.

Vitor Jorge, Universidade Federal do Rio Grande do Sul, Brazil

Enrico di Lello, Universita' degli Studi "Roma Tre", Italy

Andrea Monastero, University of Verona, Italy

Federico di Palma, University of Verona, Italy

Maura Pasquotti, Universita' degli Studi di Verona, Italy

Katalin Popovici, The MathWorks, France

Monica Reggiani, University of Padua, Italy

Maurizio di Rocco, Università degli Studi Roma Tre, Italy

P. Lopes dos Santos Santos, Instituto de Sistemas e Robótica - Porto, Portugal

Thomas Tometzki, Process Dynamics and Operations Group, Germany

Ali Emre Turgut, IRIDIA, Belgium

Maja Varga, Faculty of Electrical Engineering and Computing, Croatia

Kai Wurm, Institute of Computer Science, Germany

Selected Papers Book

A number of selected papers presented at ICINCO 2009 will be published by Springer-Verlag in a LNEE Series book. This selection will be done by the Conference chair and Program Co-chairs, among the papers actually presented at the conference, based on a rigorous review by the ICINCO 2009 Program Committee members.

FOREWORD

This book contains the proceedings of the 6th International Conference on Informatics in Control, Automation and Robotics (ICINCO 2009) which was organized by the Institute for Systems and Technologies of Information, Control and Communication (INSTICC) and held in Milan. ICINCO 2009 was co-sponsored by the International Federation for Automatic Control (IFAC) and held in cooperation with the Association for the Advancement of Artificial Intelligence (AAAI).

The ICINCO Conference Series has now consolidated as a major forum to debate technical and scientific advances presented by researchers and developers both from academia and industry, working in areas related to Control, Automation and Robotics that benefit from Information Technology.

In the Conference Program we have included oral presentations (full papers and short papers) and posters, organized in three simultaneous tracks: “Intelligent Control Systems and Optimization”, “Robotics and Automation” and “Systems Modeling, Signal Processing and Control”. We have included in the program five plenary keynote lectures, given by internationally recognized researchers, namely - Daniel S. Yeung (University of Technology, China), Maria P. Fanti (Polytechnic of Bari, Italy), Janan Zaytoon (University of Reims Champagne Ardennes, France), Alessandro Giua (Universita’ di Cagliari, Italy) and Peter S. Sapaty (Institute of Mathematical Machines and Systems, National Academy of Sciences Ukraine).

The meeting is complemented with three satellite workshops, focusing on specialized aspects of Informatics in Control, Automation and Robotics; namely, the International Workshop on Artificial Neural Networks and Intelligent Information Processing (ANNIIP), the International Workshop on Intelligent Vehicle Controls & Intelligent Transportation Systems (IVC & ITS) and the International Workshop on Networked Embedded and Control System Technologies: European and Russian R&D Cooperation (NESTER).

ICINCO received 365 paper submissions, not including those of workshops, from more than 55 countries, in all continents. To evaluate each submission, a double blind paper review was performed by the Program Committee. Finally, only 178 papers are published in these proceedings and presented at the conference. Of these, 129 papers were selected for oral presentation (34 full papers and 95 short papers) and 49 papers were selected for poster presentation. The full paper acceptance ratio was 9%, and the oral acceptance ratio (including full papers and short papers) was 35%. As in previous editions of the Conference, based on the reviewer’s evaluations and the presentations, a short list of authors will be invited to submit extended versions of their papers for a book that will be published by Springer with the best papers of ICINCO 2009.

Conferences are also meeting places where collaboration projects can emerge from social contacts amongst the participants. Therefore, in order to promote the development of research and professional networks the Conference includes in its social program a Conference and Workshops Social Event & Banquet in the evening of July 4 (Saturday).

We would like to express our thanks to all participants. First of all to the authors, whose quality work is the essence of this Conference. Next, to all the members of the Program Committee and auxiliary reviewers, who helped us with their expertise and valuable time. We would also like to deeply thank the invited speakers for their excellent contribution in sharing their knowledge and vision. Finally, a word of appreciation for the hard work of the secretariat; organizing a conference of this level is a task that can only be achieved by the collaborative effort of a dedicated and highly capable team.

Commitment to high quality standards is a major aspect of ICINCO that we will strive to maintain and reinforce next year, including the quality of the keynote lectures, of the workshops, of the papers, of the organization and other aspects of the conference. We look forward to seeing more results of R&D work in Informatics, Control, Automation and Robotics at ICINCO 2010.

Joaquim Filipe
Polytechnic Institute of Setúbal / INSTICC, Portugal

Juan Andrade Cetto
Institut de Robòtica i Informàtica Industrial CSIC-UPC, Spain

Jean-Louis Ferrier
University of Angers, France

CONTENTS

INVITED SPEAKERS

KEYNOTE SPEAKERS

SIGNAL PROCESSING, SYSTEMS MODELING AND CONTROL

FULL PAPERS

Short Papers

POSTERS

INVITED SPEAKERS

KEYNOTE SPEAKERS

SENSITIVITY BASED GENERALIZATION ERROR FOR SUPERVISED LEARNING PROBLEM WITH APPLICATIONS IN MODEL SELECTION AND FEATURE SELECTION

Daniel S. Yeung
University of Technology
China

Abstract: Generalization error model provides a theoretical support for a classifier's performance in terms of prediction accuracy. However, existing models give very loose error bounds. This explains why classification systems generally rely on experimental validation for their claims on prediction accuracy. In this talk we will revisit this problem and explore the idea of developing a new generalization error model based on the assumption that only prediction accuracy on unseen points in a neighbourhood of a training point will be considered, since it will be unreasonable to require a classifier to accurately predict unseen points "far away" from training samples. The new error model makes use of the concept of sensitivity measure for an ensemble of multiplayer feedforward neural networks (Multilayer Perceptrons or Radial Basis Function Neural Networks). Two important applications will be demonstrated, model selection and feature reduction for RBFNN classifiers. A number of experimental results using datasets such as the UCI, the 99 KDD Cup, and text categorization, will be presented.

BRIEF BIOGRAPHY

Daniel S. Yeung (Ph.D., M.Sc., M.B.A., M.S., M.A., B.A.) is the President of the IEEE Systems, Man and Cybernetics (SMC) Society, a Fellow of the IEEE and an IEEE Distinguished Lecturer. He received the Ph.D. degree in applied mathematics from Case Western Reserve University. In the past, he has worked as an Assistant Professor of Mathematics and Computer Science at Rochester Institute of Technology, as a Research Scientist in the General Electric Corporate Research Center, and as a System Integration Engineer at TRW, all in the United States. He was the chairman of the department of Computing, The Hong Kong Polytechnic University, Hong Kong, and a Chair Professor from 1999 to 2006. Currently he is a Chair Professor in the School of Computer Science and Engineering, South China University of Technology, Guangzhou, China. His current research interests include neural-network sensitivity analysis, data mining, Chinese computing, and fuzzy systems. He was the Chairman of IEEE Hong Kong Computer Chapter (91and 92), an associate editor for both IEEE Transactions on Neural Networks and IEEE Transactions on SMC (Part B), and for the International Journal on Wavelet and Multiresolution Processing. He has served as a member of the Board of Governors, Vice President for Technical Activities, and Vice President for Long Range Planning and Finance for the IEEE SMC Society. He co-founded and served as a General Co-Chair since 2002 for the International Conference on Machine Learning and Cybernetics held annually in China. He also serves as a General Co-Chair (Technical Program) of the 2006 International Conference on Pattern Recognition. He is also the founding Chairman of the IEEE SMC Hong Kong Chapter.

His past teaching and academic administrative positions include a Chair Professor and Head at Department of Computing, The Hong Kong Polytechnic University, the Head of the Management Information Unit at the Hong Kong Polytechnic University, Associate Head/Principal Lecturer at the Department of Computer Science, City Polytechnic of Hong Kong, a tenured Assistant Professor at the School of Computer Science and Technology and an Assistant Professor at the Department of Mathematics, both at Rochester Institute of Technology, Rochester, New York.

He also held industrial and business positions as a Technical Specialist/Application Software Group Leader at the Computer Consoles, Inc., Rochester, New York, an Information Resource Sub-manager/Staff Engineer at the Military and Avionics Division, TRW Inc., San Diego, California, and an Information Scientist of the Information System Operation Lab, General Electric Corporate Research and Development Centre, Schenectady, New York.

IMPACT OF THE ICT ON THE MANAGEMENT AND PERFORMANCE OF INTELLIGENT TRANSPORTATION SYSTEMS

Maria Pia Fanti
Department of Electrical and Electronic Engineering, Polytechnic of Bari, Via Re David 200, Bari, Italy
fanti@deemail.poliba.it

Keywords: Intelligent Transportation Systems, Urban Traffic Control, Intermodal Transportation Networks, Information and Communication Technologies.

Abstract: Intelligent Transportation Systems (ITS) modelling, planning, and control are research streams that, in the last years, have received a significant attention by the researcher and practitioner communities due not only to their economic impact, but also to the complexity of decisional, organizational, and management problems. Indeed, the increasing complexity of these systems and the availability of the modern ICT (Information and Communication Technologies) for the interaction among the different decision makers and for the acquisition of information by the decision makers, require both the development of suitable models and the solution of new decision problems. This presentation is aimed at showing the new attractive researches and projects in the field of ITS operational control and management in Europe. In particular, it points out the key solution of using effectively and efficiently the latest developments of ICT for ITS operational management.

1 INTRODUCTION

The term Intelligent Transportation Systems (ITS) is used to refer to technologies, infrastructure, and services, as well as the planning, operation, and control methods to be used for the transportation of person and freight. In particular, Information and Communication Technologies (ICT) are considered to be the key tools to improve efficiency and safety in transportation systems. Indeed, the advent of ICT has a tremendous impact on the planning and operations of freight transportation and on traffic management systems. ITS technologies increase the flow of available data, improve the timeliness and quality of information and offer the possibility to control and coordinate operations and traffic in real-time. Significant research efforts are required to adequately model the various planning and management problems under ITS and real-time information, and to develop efficient solution methods.

In recent years, the European Union has sponsored several projects targeting advancements of different transportation systems. On the other hand, ITS topics are considered relevant and attractive research areas.

In Section 2 the paper recalls the most important European Projects in the fields of ITS and intelligent freight transportation. Moreover, Sections 3 and 4 present the research advances in two crucial sectors of ITS: the management of Urban Traffic Networks (UTN) and of Intermodal Transportation Networks (ITN), respectively.

2 EUROPEAN PROJECTS IN THE FIELD OF ITS

A basic project on ITS is CESAR I & II (Co-operative European System for Advanced Information Redistribution) that proposes an Internet communication platform that aims to integrate services and data for unaccompanied traffic and the rolling motorway traffic management. Moreover, in the field of railway system management, CroBIT (Cross Border Information Technology) is a new system that provides the railways with a tool to track consignments and integrates freight railways along a transport corridor providing total shipment visibility. A maritime navigation information structure in

European waters is established by MarNIS (Maritime Navigation Information Systems) that is an integrated project aiming to develop tools that can be used to exchange maritime navigation information and to improve safety, security and efficiency of maritime traffic.

In addition, several projects focus specifically on efficient freight transportation. For instance, Freightwise aims to establish a framework for efficient co-modal freight transport on the Norwegian ARKTRANS system. One of the main objectives in Freightwise is establishing a framework for efficient co-modal freight transport and simplifying the interaction among stakeholders during planning, execution and completion of transport operations. Moreover, the project e-Freight is a continuation of Freightwise to promote efficient and simplified solutions in support of cooperation, interoperability and consistency in the European Transport System. E-Freight is to support the Freight Logistics Action Plan, which focuses on quality and efficiency for the movement of goods, as well as on ensuring that freight-related information travels easily among modes. Furthermore, in the Seventh Framework Program (FP7-ICT Objective 6.1), the SMARTFREIGHT project wants to make urban freight transport more efficient, environmentally friendly and safe by answering to challenges related to traffic management and the relative coordination. Indeed, freight distribution management in city centres is usually operated by several commercial companies and there is no coordination of these activities in a way that would benefit the city. The main aim of SMARTFREIGHT is therefore to specify, implement and evaluate ICT solutions that integrate urban traffic management systems with the management of freight and logistics in urban areas. Finally, EURIDICE (European Inter-Disciplinary Research on Intelligent Cargo for Efficient, Safe and Environment-friendly Logistics) is a project sponsored by the European Commission under the 7th Framework Program seeking to develop an advanced European logistics system around the concept of ‘intelligent cargo’. The goal is networking cargo objects like packages, vehicles and containers to provide information services whenever required along the transport chain. The project aims to build an information service platform centred on the individual cargo item and its interaction with the surrounding environment and the user.

3 URBAN TRAFFIC MANAGEMENT

Traffic congestion of urban roads undermines mobility in major cities. Traditionally, the congestion problem on surface streets was dealt by adding more lanes and new links to the existing Urban Traffic Networks (UTN). Since such a solution can no longer be considered for limited availability of space in urban centres, greater emphasis is nowadays placed on traffic management through the implementation and operation of ICT. In particular, traffic signal control on surface street networks plays a central role in traffic management. Despite the large research efforts on the topic, the problem of urban intersection congestion remains an open issue (Lo, 2001, Papageorgiou, 1999). Most of the currently implemented traffic control systems may be grouped into two principal classes (Papageorgiou et al., 2003, Patel and Ranganathan, 2001): i) fixed time strategies, that are derived off-line by use of optimization codes based on historical traffic data; ii) vehicle actuated strategies, that perform an on-line optimization and synchronization of the signal timing plans and make use of real time measurements. While the fixed time strategies do not use information on the actual traffic situation, the second actuated control class can be viewed as a traffic responsive network signal policy employing signal timing plans that respond automatically to traffic conditions. In a real time control strategy, detectors located on the intersection approaches monitor traffic conditions and feed information on the actual system state to the real time controller, which selects the duration of the phases in the signal timing plan in order to optimize an objective function. Although the corresponding optimal control problem may readily be formulated, its real time solution and realization in a control loop has to face several difficulties such as the size and the combinatorial nature of the optimization problem, the measurements of traffic conditions and the presence of unpredictable disturbances. The first and most notable of vehicle actuated techniques is the British SCOOT (Hunt et al., 1982), that decides an incremental change of splits, offsets and cycle times based on real time measurements. However, although SCOOT exhibits a centralized hardware architecture, the strategy is functionally decentralized with regard to splits setting. A formulation of the traffic signal network optimization strategy is presented in (Lo, 2001) and (Wey, 2000). However, the resulting procedures lead to complex mixed integer linear programming problems that are computationally intensive and the formulation for real networks requires heuristics for solutions. Furthermore, Diakaki *et al.* (2002) propose a traffic responsive urban control strategy based on a feedback approach involving the application of a systematic and powerful control design method. Despite the simplicity and the efficiency of the proposed control strategy, such a modelling approach can not directly consider the effects of offset for consecutive junctions and the time-variance of the turning rates and the saturation flows.

An improvement on urban traffic actuated control strategies is provided in (Dotoli et al., 2006) where the green splits for a fixed cycle time are determined in real time, in order to minimize the number of vehicles in queue in the considered signalized area. The paper gives a contribution in facing the *apparently insurmountable difficulties* (Papageorgiou et al., 2003) in the real time solution and realization of the control loop governing an urban intersection by traffic lights. To this aim, the paper pursues simplicity in the modelling and in the optimization procedure by presenting a macroscopic model to describe the urban traffic network. Describing the system by a discrete time model with the sampling time equal to the cycle, the timing plan is obtained on the basis of the real traffic knowledge and the traffic measurements in a prefixed set of cycles. The traffic urban control strategy is performed by solving a mathematical programming problem that minimizes the number of vehicles in the considered urban area. The minimization of the objective function is subject to linear constraints derived from the intersection topology, the fixed cycle duration and the minimum and maximum duration of the phases commonly adopted in practice. The optimization problem is solved by a standard optimization software on a personal computer, so that practical applications are possible in a real time control framework.

4 INTERMODAL TRANSPORTATION NETWORKS

Intermodal Transportation Networks (ITN) are logistics systems integrating different transportation services, designed to move goods from origin to destination in a timely manner and using multiple modes of transportation (rail, ocean vessel, truck etc.). In the related literature several papers analyze ITN operations and planning issues as container fleet management, container terminal operations and scheduling. With the development of the new ICT tools, these operative and planning issues can be dealt with in a different way. In fact, these new technologies can effectively impact on the planning and operation of ITS. In particular, ICT solutions can increase the data flow and the information quality while allowing real-time data exchange in transportation systems (Crainic and Kim, 2007, Ramstedt and Woxenius, 2006). As mentioned in (Giannopoulos, 2004), numerous new applications of ICT to the transportation field are in various stages of development, but in the information transfer area the new systems seem to be too unimodal. In the application of ICT solutions to multimodal chains, an important and largely unexplored research field is the assessment of the impact of new technologies before their implementation, by a cost-benefit analysis (Zografos and Regan, 2004, Crainic and Kim, 2007). This research field offers numerous research opportunities: for instance, a not well explored case is that of coordinating independent stakeholders in the presence of uncertainties and lack of information on the stakeholders operations and their propagation within the intermodal chain.

An efficient ITN needs to synchronize the logistics operations. Therefore, information exchange among stakeholders is essential and ICT solutions are key tools to achieve efficiency. Nevertheless, the increasing complexity of these systems and the availability of the modern ICT for the interaction among the different decision makers and for the acquisition of information by the decision makers, require both the development of suitable models and the solution of new decision problems. Moreover, ITN and their decision making process are complex systems characterized by a high degree of interaction, concurrency and synchronization. Hence, ITS can be modeled as Discrete Event Systems (DES), whose dynamics depends on the interaction of discrete events, such as demands, departures and arrivals of means of transport at terminals and acquisitions and releases of resources by vehicles. DES models are widely used to describe decision making and operational processes. In the domain of ITN, the potentialities of these models are not fully explored and exploited. In particular at the operational level, we recall the models in the Petri net (Peterson, 1981) frameworks (Danielis et al., 2009, Di Febbraro et al., 2006, Fischer et al., 2000) and the simulation models (Boschian et al., 2009, Parola and Sciomachen, 2005).

In this presentation we mention two papers (Boschian et al., 2009) and (Danielis et al., 2009) that point out the role and the impact of the ICT applications in the field of the ITN management and control. In particular, paper (Danielis et al., 2009) focuses on the ICT solutions that allow sharing information among stakeholders on the basis of user friendly technologies. To this aim the authors single out some performance indices to evaluate activities, resources (utilization) and output (throughput, lead time) by integrating information flows allowed by the use of ICT tools. A case study is analyzed considering an ITN constituted by a port and a truck terminal of an Italian town including the road-ship transshipment process. The system is modeled and simulated in a timed Petri net framework considering different dynamic conditions characterized by a diverse level of information shared between terminals and operators. The simulation results show that ICT have a huge potential for efficient real time management and operation of ITN, as well as an effective impact on the infrastructures, reducing both the utilization of the system resources as well as the cost performance indices.

An application of the ICT tools to the real-time transport monitoring in order to trace and safely handle moving goods is presented in (Boschian et al., 2009). In particular, the authors analyze and simulate a real case study involving an ITN system and the transport and the customs clearance of goods that arrive to the port and the intermodal terminal. The case study is analyzed in the frame of the EURIDICE Integrated Project. The structure and the dynamics of the ITN model is described by the Unified Modeling Language formalism (Miles and Hamilton, 2006) and is implemented by a discrete-event simulation in Arena environment. The task is to provide services for the efficient utilization of infrastructures, both singularly and across territorial networks (e.g., port terminal synchronization with rail and road connections) and to contain the impact of logistic infrastructures on the local communities, reducing congestion and pollution caused by the associated freight movements. The discrete event simulation study shows that the application of the ICT tools allows us to locate goods and the related up-to-date information and to extend it with useful information-based services. Summing up, the simulation results show that integrating ICT into the system leads to a more efficient system management and drastically reduces the system lead times.

5 CONCLUSIONS

The paper presents the new attractive researches and projects in the field of ITS operational control and management. In particular, the key solutions of using effectively and efficiently the latest developments of ICT for ITS operational management are pointed out. The presentation focuses on the most important European Projects in ITS and on two crucial fields of the ITS management and control: the management of Urban Traffic networks and of Intermodal Transportation Networks. In the two cases are emphasized the new results and the challenges of future researches.

REFERENCES

Boschian, V., Fanti, M.P., Iacobellis, G., Ukovich, W., 2009. Using Information and Communication Technologies in Intermodal Freight Transportation: a Case Study. Submitted for publication.

Crainic, T.G., Kim, K.H., 2007 Intermodal transportation, In: C. Barnhart and G. Laporte, Editors, *Transportation, Handbooks in Operations Research and Management Science*, vol. 14, North-Holland, Amsterdam, pp. 467–537.

Danielis, R., Dotoli, M., Fanti, M.P., Mangini, A.M., Pesenti R., Stecco G., Ukovich W., 2009, “Integrating ICT into Logistics Intermodal Systems: A Petri Net Model of the Trieste Port”, The European Control Conference 2009, ECC’09, August 23-26, Budapest, Hungary.

Diakaki, C., Papageorgiou, M., Aboudolas, K. 2002. A multivariable regulator approach to traffic-responsive network-wide signal control. *Control Engineering Practice*, *10*(2), 183-195.

Di Febbraro, A., Giglio, D., Sacco, N., 2002. On applying Petri nets to determine optimal offsets for coordinated traffic light timings. *Proc. 5th IEEE Int. Conf. on Intelligent Transportation Systems* (pp. 773-778), Singapore.

Di Febbraro, A., Giglio, D., Sacco, N., 2004. Urban traffic control structure based on hybrid Petri nets. *IEEE Trans. On Intelligent Transportation Systems 5,* (4), 224-237.

Di Febbraro, A., Porta, G., N. Sacco, N., 2006. A Petri net modelling approach of intermodal terminals based on Metrocargo system, *Proc. Intelligent Transportation Systems Conf.*, pp. 1442–1447.

Dotoli, M., Fanti, M.P., Meloni, C., 2006. A Signal Timing Plan Formulation for Urban Traffic Control. *Control Engineering Practice*, vol. 14, no.11, 2006, pp. 1297-1311.

Fischer, M., Kemper, P., 2000. Modeling and analysis of a freight terminal with stochastic Petri nets. In *Proc. 9th IFAC Symposium Control in Transportation Systems*.

Giannopoulos, G.A., 2004. The application of information and communication technologies in transport. In *European Journal Of Operational Research*, vol. 152, pp. 302-320.

Hunt, P.B., Robertson, D.L., Beterton, R.D., & Royle, M.C., 1982. The SCOOT on-line traffic signal optimization technique. *Traffic Engineering and Control, 23,* 190-199.

Lo, H. K., 2001. A cell-based traffic control formulation: strategies and benefits of dynamic timing plans. *Transportation Science*, *35*(2), 148-164.

Miles, R., Hamilton, K., 2006. *Learning UML 2.0.* O'Reilly Media, Sabastopol CA USA.

Papageorgiou, M., 1999. Automatic control methods in traffic and transportation. In *Operations research and decision aid methodologies in traffic and transportation management*, P. Toint, M. Labbe, K. Tanczos, & G. Laporte (Eds.), Springer-Verlag, 46-83.

Papageorgiou, M., Diakaki, C., Dinopoulou, V., Kotsialos, A., Wang, Y., 2003. Review of road traffic control strategies. *Proceedings of the IEEE*, *91*(12), 2043-2067.

Parola F., Sciomachen, A., 2005. Intermodal container flows in a port system metwor: Analysis of possible growths via simulation models. In *International Journal of Production Economics*, vol. 97, pp. 75-88.

Patel, M., Ranganathan, N., 2001. IDUTC: an intelligent decision-making system for urban traffic control applications. In *IEEE Trans. on Vehicular Technology*, *50*(3), 816-829.

Peterson, J.L., 1981. *Petri Net Theory and the Modeling of Systems*. Prentice Hall, Englewood Cliffs, NJ, USA.

Ramstedt, L., Woxenius, J., 2006. Modelling approaches to operational decision-making in freight transport chains, *Proc. 18th NOFOMA Conf.*.

Wey W.-M., 2000. Model formulation and solution algorithm of traffic signal control in an urban network. In *Computers, Environment and Urban Systems*, *24*(4), 355-377.

Zografos, K.G., Regan, A., 2004. Current Challenges for Intermodal Freight Transport and Logistics in Europe and the US. In *Journal of the Transportation Research Board*, vol. 1873, pp. 70-78.

BRIEF BIOGRAPHY

Maria Pia Fanti is associate professor in Systems and Control Engineering and is with the Department of Electrical and Electronic Engineering of the Polytechnic of Bari (Italy). Maria Pia Fanti received the Laurea degree in Electronic Engineering from the University of Pisa (Italy), in 1983 and obtained an IBM thesis award. She was a visiting researcher at the Rensselaer Polytechnic Institute of Troy, New York, in 1999. Her research interests include discrete event systems, Petri nets, modeling and control of automated manufacturing systems, modeling and management of logistics system, supply chains and health care systems.

Prof. Fanti is Associate Editor of the following journals: IEEE Trans. on Systems, Man, and Cybernetics. Part A, IEEE Trans. on Automation Science and Engineering, The Mediterranean J. of Measurement and Control, Int. J. of Automation and Control, and Enterprise Information Systems. She is Co-Chair of the Technical committee on Discrete Event Systems for the IEEE SMC Society, Chair of the Italy Section SMC Chapter, and member of the IFAC Technical Committee on Discrete Event and Hybrid Systems. She is authors of 120+ papers. She has served in 20+ conference international program committees, she is IPC chair of 2nd IFAC Workshop on Dependable Control of Discrete Systems, Bari, Italy, 2009 and of the IEEE Workshop on Health Care Management, Venice, Italy, 2010.

RECENT ADVANCES IN VERIFICATION AND ANALYSIS OF HYBRID SYSTEMS

Janan Zaytoon
University of Reims Champagne Ardennes
France

Abstract: Formal verification of properties is a very important area of analysis of hybrid systems. It is, indeed, essential to use methods and tools to guarantee that the global behaviour of a system is correct and consistent with the specifications. This is especially true for safety properties that insure that the system is not dangerous for itself or its environment.

Classically, verification of Safety properties may be performed with reachability computation in the hybrid state space. Basic ideas have not really evolved since the first works, however new techniques have been proposed and algorithms have been improved.

The aim of this talk is to present the problem of verification and reachability computation for hybrid systems and to propose a classification of recent improvements. To overcome the difficulties in verification and reachability analysis it is necessary to make choices regarding general principles, algorithms and mathematical representation of regions of the continuous state space. These choices depend on each other and must be consistent. However all approaches are based on common considerations that will be used to structure the talk.

BRIEF BIOGRAPHY

Born in 1962, Janan Zaytoon (BSc Eng./1983, MSc Eng./1986, DEA/1988, PhD/1993, Habilitation/1997) is Professor and Head of the CReSTIC Research Centre (involving 150 researchers) at Reims University. He was a member of the Administration Council of the same University (2003-2006). He is the Chair of the French national research network/group "GDR MACS of CNRS", which involves all the researchers in the field of Automatic Control Syetems in France (about 2000 researchers and PhD students).

His involvement in IFAC includes his service as member of the IFAC Council from 2008 to 2011, head of the French National Member Organizer since 1999, Chair of Technical Committee on Discrete Event and Hybrid Systems from 2005 to 2008, Vice-Chair of this Technical Committee from 2002 to 2005 and 2008 to 2011, member of the Publication Committee of IFAC from 2008 to 2011, Editor of the IFAC Journal "Control Engineering Practice" and the Affiliated IFAC Journal "Nonlinear Analysis: Hybrid Systems".

Professor Janan Zaytoon is the author/co-author of 70 journal papers, 3 books, 12 book chapters, 120 conference papers, and 8 patents. His main research interests are in the fields of Discrete Event Systems, Hybrid Dynamic Systems, Intelligent Control and Biomedical Engineering. He is an associate Editor of "IET Control Theory and Applications" and "Discrete Event Dynamic Systems", IPC and/or NOC Chair/Co-Chair of 15 Conferences, Editor/Co-Editor of 10 Conference Proceedings, Keynote speaker for 6 conferences, supervisor of 20 PhD students, Guest Editor/Co-editor for 18 special issues of 6 international and 2 national journals, leader of 8 industrial contracts, and was Chair of the WODES (International Workshop on Discrete Event Systems) steering Committee.

DISCRETE EVENT DIAGNOSIS USING PETRI NETS

Maria Paola Cabasino, Alessandro Giua and Carla Seatzu
Department of Electrical and Electronic Engineering, University of Cagliari, Piazza D'Armi, 09123 Cagliari, Italy
{*cabasino, giua, seatzu*}*@diee.unica.it*

Keywords: Petri nets, Diagnosis, Discrete event systems.

Abstract: This paper serves as a support for the plenary address given by the second author during the conference. In this paper we present an approach to on-line diagnosis of discrete event systems based on labeled Petri nets, that are a particular class of Petri nets where some events are undistinguishable, i.e., events that produce an output signal that is observable, but that is common to other events. Our approach is based on the notion of basis markings and justifications and it can be applied both to bounded and unbounded Petri nets whose unobservable subnet is acyclic. Moreover it is shown that, in the case of bounded Petri nets, the most burdensome part of the procedure may be moved off-line, computing a particular graph that we call *Basis Reachability Graph*.

Finally we present a diagnosis MATLAB toolbox with some examples of application.

1 INTRODUCTION

Failure detection and isolation in industrial systems is a subject that has received a lot of attention in the past few decades. A failure is defined to be any deviation of a system from its normal or intended behavior. Diagnosis is the process of detecting an abnormality in the system behavior and isolating the cause or the source of this abnormality.

Failures are inevitable in today's complex industrial environment and they could arise from several sources such as design errors, equipment malfunctions, operator mistakes, and so on. As technology advances, as we continue to build systems of increasing size and functionality, and as we continue to place increasing demands on the performance of these systems, then so do we increase the complexity of these systems. Consequently (and unfortunately), we enhance the potential for systems to fail, and no matter how safe our designs are, how improved our quality control techniques are, and how better trained the operators are, system failures become unavoidable.

Given the fact that failures are inevitable, the need for effective means of detecting them is quite apparent if we consider their consequences and impacts not just on the systems involved but on the society as a whole. Moreover we note that effective methods of failure diagnosis can not only help avoid the undesirable effects of failures, but can also enhance the operational goals of industries. Improved quality of performance, product integrity and reliability, and reduced cost of equipment maintenance and service are some major benefits that accurate diagnosis schemes can provide, especially for service and product oriented industries such as home and building environment control, office automation, automobile manufacturing, and semiconductor manufacturing. Thus, we see that accurate and timely methods of failure diagnosis can enhance the safety, reliability, availability, quality, and economy of industrial processes.

The need of automated mechanisms for the timely and accurate diagnosis of failures is well understood and appreciated both in industry and in academia. A great deal of research effort has been and is being spent in the design and development of automated diagnostic systems, and a variety of schemes, differing both in their theoretical framework and in their design and implementation philosophy, have been proposed.

In diagnosis approach two different problems can be solved: the problem of diagnosis and the problem of diagnosability.

Solving a problem of diagnosis means that we associate to each observed string of events a diagnosis state, such as “normal” or “faulty” or “uncertain”. Solving a problem of diagnosability is equivalent to determine if the system is diagnosable, i.e., to determine if, once a fault has occurred, the system can detect its occurrence in a finite number of steps.

The diagnosis of discrete event systems (DES) is a research area that has received a lot of attention in the last years and has been motivated by the practical need of ensuring the correct and safe functioning of large complex systems. As discussed in the next session the first results have been presented within the framework of automata. More recently, the diagnosis problem has also been addressed using Petri nets (PNs). In fact, the use of Petri nets offers significant advantages because of their twofold representation: graphical and mathematical. Moreover, the intrinsically distributed nature of PNs where the notion of state (i.e., marking) and action (i.e., transition) is local reduces the computational complexity involved in solving a diagnosis problem.

In this paper we summarize our main contributions on diagnosis of DES using PNs (Giua and Seatzu, 2005; Cabasino et al., 2008; Lai et al., 2008; Cabasino et al., 2009). In particular, we focus on arbitrary labeled PNs where the observable events are the labels associated to transitions, while faults are modeled as silent transitions. We assume that there may also be transitions modeling a regular behavior, that are silent as well. Moreover, two or more transitions that may be simultaneously enabled may share the same label, thus they are undistinguishable. Our diagnosis approach is based on the definition of four diagnosis states modeling different degrees of alarm and it applies to all systems whose unobservable subnet is acyclic. Two are the main advantages of our procedure. First, we do not need an exhaustive enumeration of the states in which the system may be: this is due to the introduction of basis markings. Secondly, in the case of bounded net systems we can move off-line the most burdensome part of the procedure building a finite graph called basis reachability graph.

The paper is organized as follows. In Section 2 the state of art of diagnosis for discrete event systems is illustrated. In Section 3 we provide a background on PNs. In Sections 4 and 5 are introduced the definitions of minimal explanations, justifications and basis markings, that are the basic notions of our diagnosis approach. In Section 6 the diagnosis states are defined and a characterization of them in terms of basis markings and j-vectors is given. In Section 7 we show how the most burdensome part of the procedure can be moved offline in the case of bounded PNs. In Section 8 we present the MATLAB toolbox developed by our group for PNs diagnosis and in Section 9 we present some numerical results obtained applying our tool to a parametric model of manufacturing system. In Section 10 we draw the conclusions.

2 LITERATURE REVIEW

In this section we present the state of art of diagnosis of DES using automata and PNs.

2.1 Diagnosis of DES using Automata

In the contest of DES several original theoretical approaches have been proposed using *automata*.

In (Lin, 1994) and (Lin et al., 1993) a state-based DES approach to failure diagnosis is proposed. The problems of off-line and on-line diagnosis are addressed separately and notions of diagnosability in both of these cases are presented. The authors give an algorithm for computing a diagnostic control, i.e., a sequence of test commands for diagnosing system failures. This algorithm is guaranteed to converge if the system satisfies the conditions for on-line diagnosability.

In (Sampath et al., 1995) and (Sampath et al., 1996) the authors propose an approach to failure diagnosis where the system is modeled as a DES in which the failures are treated as unobservable events. The level of detail in a discrete event model appears to be quite adequate for a large class of systems and for a wide variety of failures to be diagnosed. The approach is applicable whenever failures cause a distinct change in the system status but do not necessarily bring the system to a halt. In (Sampath et al., 1995) a definition of diagnosability in the framework of formal languages is provided and necessary and sufficient conditions for diagnosability of systems are established. Also presented in (Sampath et al., 1995) is a systematic approach to solve the problem of diagnosis using diagnosers.

In (Sampath et al., 1998) the authors present an integrated approach to control and diagnosis. More specifically, authors present an approach for the design of diagnosable systems by appropriate design of the system controller and this approach is called active diagnosis. They formulate the active diagnosis problem as a supervisory control problem. The adopted procedure for solving the active diagnosis problem is the following: given the non-diagnosable language generated by the system of interest, they first select an “appropriate” sublanguage of this language as the legal language. Choice of the legal language is a design issue and typically depends on considerations such as acceptable system behavior (which ensures that the system behavior is not restricted more than necessary in order to eventually make it diagnosable) and detection delay for the failures. Once the appropriate legal language is chosen, they then design a controller (diagnostic controller), that achieves a closed-loop language that is within the legal language and is diagnosable. This controller is designed based on the formal framework and the synthesis techniques that supervisory control theory provides, with the additional constraint of diagnosability.

In (Debouk et al., 2000) is addressed the problem of failure diagnosis in DES with decentralized information. Debouk *et al.* propose a coordinated decentralized architecture consisting of two local sites communicating with a coordinator that is responsible for diagnosing the failures occurring in the system. They extend the notion of diagnosability, originally introduced in (Sampath et al., 1995) for centralized systems, to the proposed coordinated decentralized architecture. In particular, they specify three protocols that realize the proposed architecture and analyze the diagnostic properties of these protocols.

In (Boel and van Schuppen, 2002) the authors address the problem of synthesizing communication protocols and failure diagnosis algorithms for decentralized failure diagnosis of DES with costly communication between diagnosers. The costs on the communication channels may be described in terms of bits and complexity. The costs of communication and computation force the trade-off between the control objective of failure diagnosis and that of minimization of the costs of communication and computation. The results of this paper is an algorithm for decentralized failure diagnosis of DES for the special case of only two diagnosers.

In (Zad et al., 2003) a state-based approach for online passive fault diagnosis is presented. In this framework, the system and the diagnoser (the fault detection system) do not have to be initialized at the same time. Furthermore, no information about the state or even the condition (failure status) of the system before the initiation of diagnosis is required. The design of the fault detection system, in the worst case, has exponential complexity. A model reduction scheme with polynomial time complexity is introduced to reduce the computational complexity of the design. Diagnosability of failures is studied, and necessary and sufficient conditions for failure diagnosability are derived.

2.2 Diagnosis of DES using Petri Nets

Among the first pioneer works dealing with PNs, we recall the approach of Prock. In (Prock, 1991) the author proposes an on-line technique for fault detection that is based on monitoring the number of tokens residing into P-invariants: when the number of tokens inside P-invariants changes, then the error is detected.

In (Sreenivas and Jafari, 1993) the authors employ time PNs to model the DES controller and backfiring transitions to determine whether a given state is invalid. Later on, time PNs have been employed in (Ghazel et al., 2005) to propose a monitoring approach for DES with unobservable events and to represent the “a priori” known behavior of the system, and track on-line its state to identify the events that occur.

In (Hadjicostis and Veghese, 1999) the authors use PN models to introduce redundancy into the system and additional P-invariants allow the detection and isolation of faulty markings.

Redundancy into a given PN is used in (Wu and Hadjicostis, 2005) to enable fault detection and identification using algebraic decoding techniques. In this paper Wu and Hadjicostis consider two types of faults: place faults that corrupt the net marking, and transition faults that cause a not correct update of the marking after event occurrence. Although this approach is general, the net marking has to be periodically observable even if unobservable events occur. Analogously, in (Lefebvre and Delherm, 2007) the authors investigate on the determination of the set of places that must be observed for the exact and immediate estimation of faults occurrence.

In (Ruiz-Beltràn et al., 2007) Interpreted PNs are employed to model the system behavior that includes both events and states partially observable. Based on the Interpreted PN model derived from an on-line methodology, a scheme utilizing a solution of a programming problem is proposed to solve the problem of diagnosis.

Note that, all papers in this topic assume that faults are modeled by unobservable transitions. However, while the above mentioned papers assume that the marking of certain places may be observed, a series of papers have been recently presented that are based on the assumption that no place is observable (Basile et al., 2008; Benveniste et al., 2003; Dotoli et al., 2008; Genc and Lafortune, 2007).

In particular, in (Genc and Lafortune, 2007) the authors propose a diagnoser on the basis of a modular approach that performs the diagnosis of faults in each module. Subsequently, the diagnosers recover the monolithic diagnosis information obtained when all the modules are combined into a single module that preserves the behavior of the underlying modular system. A communication system connects the different modules and updates the diagnosis information. Even if the approach does not avoid the state explosion problem, an improvement is obtained when the system can be modeled as a collection of PN modules coupled through common places.

The main advantage of the approaches in (Genc and Lafortune, 2007) consists in the fact that, if the net is bounded, the diagnoser may be constructed off-line, thus moving off-line the most burdensome part of the procedure. Nevertheless, a characterization of the set of markings consistent with the actual observation is needed. Thus, large memory may be required.

An improvement in this respect has been given in (Benveniste et al., 2003; Basile et al., 2008; Dotoli et al., 2008).

In particular, in (Benveniste et al., 2003) a net unfolding approach for designing an on-line asynchronous diagnoser is used. The state explosion is avoided but the on-line computation can be high due to the on-line building of the PN structures by means of the unfolding.

In (Basile et al., 2008) the diagnoser is built on-line by defining and solving Integer Linear Programming (ILP) problems. Assuming that the fault transitions are not observable, the net marking is computed by the state equation and, if the marking has negative components, an unobservable sequence is occurred. The linear programming solution provides the sequence and detects the fault occurrences. Moreover, an off-line analysis of the PN structure reduces the computational complexity of the ILP problem.

In (Dotoli et al., 2008), in order to avoid the redesign and the redefinition of the diagnoser when the structure of the system changes, the authors propose a diagnoser that works on-line. In particular, it waits for an observable event and an algorithm decides whether the system behavior is normal or may exhibit some possible faults. To this aim, some ILP problems are defined and provide eventually the minimal sequences of unobservable transitions containing the faults that may have occurred. The proposed approach is a general technique since no assumption is imposed on the reachable state set that can be unlimited, and only few properties must be fulfilled by the structure of the PN modeling the system fault behavior.

We also proposed a series of contributions dealing with diagnosis of PNs (Giua and Seatzu, 2005; Cabasino et al., 2008; Lai et al., 2008; Cabasino et al., 2009). Our main results are summarized in the rest of the paper.

Note that none of the above mentioned papers regarding PNs deal with *diagnosability*, namely none of them provide a procedure to determine a priori if a system is *diagnosable*, i.e., if it is possible to reconstruct the occurrence of fault events observing words of finite length.

In fact, whereas this problem has been extensively studied within the framework of automata as discussed above, in the PN framework very few results have been presented.

The first contribution on diagnosability of PNs was given in (Ushio et al., 1998). They extend a necessary and sufficient condition for diagnosability in (Sampath et al., 1995; Sampath et al., 1996) to unbounded PN. They assume that the set of places is partitioned into observable and unobservable places, while all transitions are unobservable in the sense that their occurrences cannot be observed. Starting from the PN they build a diagnoser called *simple* ω *diagnoser* that gives them sufficient conditions for diagnosability of unbounded PNs.

In (Chung, 2005) the authors, in contrast with Ushio's paper, assumes that part of the transitions of the PN modelling is observable and shows as the additional information from observed transitions in general adds diagnosability to the analysed system. Moreover starting from the diagnoser he proposes an automaton called *verifier* that allows a polynomial check mechanism on diagnosability but for finite state automata models.

In (Wen and Jeng, 2005) the authors propose an approach to test diagnosability by checking the structure property of T-invariant of the nets. They use Ushio's diagnoser to prove that their method is correct, however they don't construct a diagnoser for the system to do diagnosis. In (Wen et al., 2005) they also present an algorithm, based on a linear programming problem, of polynomial complexity in the number of nodes for computing a sufficient condition of diagnosability of DES modeled by PN.

3 BACKGROUND

In this section we recall the formalism used in the paper. For more details on PNs we refer to (Murata, 1989).

A *Place/Transition net* (P/T net) is a structure $N = (P, T, Pre, Post)$, where P is a set of m places; T is a set of n transitions; $Pre : P \times T \to \mathbb{N}$ and $Post : P \times T \to \mathbb{N}$ are the *pre*– and *post*– incidence functions that specify the arcs; $C = Post - Pre$ is the incidence matrix.

A *marking* is a vector $M : P \to \mathbb{N}$ that assigns to each place of a P/T net a non–negative integer number of tokens, represented by black dots. We denote $M(p)$ the marking of place p. A P/T *system* or *net system* $\langle N, M_0 \rangle$ is a net N with an initial marking M_0. A transition t is enabled at M iff $M \geq Pre(\cdot, t)$ and may fire yielding the marking $M' = M + C(\cdot, t)$. We write $M\,[\sigma\rangle$ to denote that the sequence of transitions $\sigma = t_{j_1} \cdots t_{j_k}$ is enabled at M, and we write $M\,[\sigma\rangle\,M'$ to denote that the firing of σ yields M'. We also write $t \in \sigma$ to denote that a transition t is contained in σ.

The set of all sequences that are enabled at the initial marking M_0 is denoted $L(N,M_0)$, i.e., $L(N,M_0) = \{\sigma \in T^* \mid M_0[\sigma\rangle\}$.

Given a sequence $\sigma \in T^*$, we call $\pi : T^* \to \mathbb{N}^n$ the function that associates to σ a vector $y \in \mathbb{N}^n$, named the *firing vector* of σ. In particular, $y = \pi(\sigma)$ is such that $y(t) = k$ if the transition t is contained k times in σ.

A marking M is *reachable* in $\langle N, M_0\rangle$ iff there exists a firing sequence σ such that $M_0\ [\sigma\rangle\ M$. The set of all markings reachable from M_0 defines the *reachability set* of $\langle N, M_0\rangle$ and is denoted $R(N,M_0)$.

A PN having no directed circuits is called *acyclic*. A net system $\langle N, M_0\rangle$ is *bounded* if there exists a positive constant k such that, for $M \in R(N,M_0)$, $M(p) \leq k$.

A *labeling function* $\mathcal{L} : T \to L \cup \{\varepsilon\}$ assigns to each transition $t \in T$ either a symbol from a given alphabet L or the empty string ε.

We denote as T_u the set of transitions whose label is ε, i.e., $T_u = \{t \in T \mid \mathcal{L}(t) = \varepsilon\}$. Transitions in T_u are called *unobservable* or *silent*. We denote as T_o the set of transitions labeled with a symbol in L. Transitions in T_o are called *observable* because when they fire their label can be observed. Note that in this paper we assume that the same label $l \in L$ can be associated to more than one transition. In particular, two transitions $t_1, t_2 \in T_o$ are called *undistinguishable* if they share the same label, i.e., $\mathcal{L}(t_1) = \mathcal{L}(t_2)$. The set of transitions sharing the same label l are denoted as T_l.

In the following we denote as C_u (C_o) the restriction of the incidence matrix to T_u (T_o) and denote as n_u and n_o, respectively, the cardinality of the above sets. Moreover, given a sequence $\sigma \in T^*$, $P_u(\sigma)$, resp., $P_o(\sigma)$, denotes the projection of σ over T_u, resp., T_o.

We denote as w the word of events associated to the sequence σ, i.e., $w = P_o(\sigma)$. Note that the length of a sequence σ (denoted $|\sigma|$) is always greater than or equal to the length of the corresponding word w (denoted $|w|$). In fact, if σ contains k' transitions in T_u then $|\sigma| = k' + |w|$.

Definition 3.1 (Cabasino et al., 2009). Let $\langle N, M_0\rangle$ be a labeled net system with labeling function $\mathcal{L} : T \to L \cup \{\varepsilon\}$, where $N = (P, T, Pre, Post)$ and $T = T_o \cup T_u$. Let $w \in L^*$ be an observed word. We define

$$\mathcal{S}(w) = \{\sigma \in L(N,M_0) \mid P_o(\sigma) = w\}$$

the set of firing sequences *consistent* with $w \in L^*$, and

$$\mathcal{C}(w) = \{M \in \mathbb{N}^m \mid \exists \sigma \in T^* : P_o(\sigma) = w \wedge M_0[\sigma\rangle M\}$$

the set of markings *consistent* with $w \in L^*$. ■

In plain words, given an observation w, $\mathcal{S}(w)$ is the set of sequences that may have fired, while $\mathcal{C}(w)$ is the set of markings in which the system may actually be.

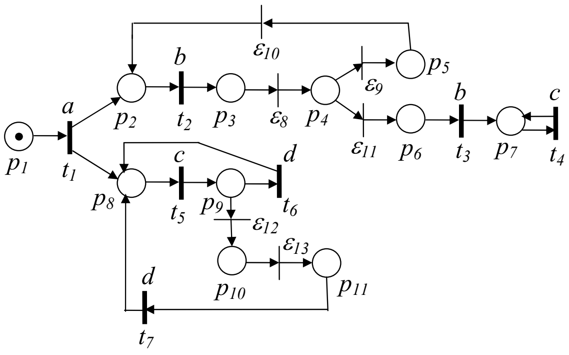


Figure 1: A PN system modeling.

Example 3.2. Let us consider the PN in Figure 1. Let us assume $T_o = \{t_1, t_2, t_3, t_4, t_5, t_6, t_7\}$ and $T_u = \{\varepsilon_8, \varepsilon_9, \varepsilon_{10},\ \varepsilon_{11}, \varepsilon_{12}, \varepsilon_{13}\}$, where for a better understanding unobservable transitions have been denoted ε_i rather than t_i. The labeling function is defined as follows: $\mathcal{L}(t_1) = a$, $\mathcal{L}(t_2) = \mathcal{L}(t_3) = b$, $\mathcal{L}(t_4) = \mathcal{L}(t_5) = c$, $\mathcal{L}(t_6) = \mathcal{L}(t_7) = d$.

First let us consider $w = ab$. The set of firing sequences that is consistent with w is $\mathcal{S}(w) = \{t_1t_2,\ t_1t_2\varepsilon_8, t_1t_2\varepsilon_8\varepsilon_9, t_1t_2\varepsilon_8\varepsilon_9\varepsilon_{10}, t_1t_2\varepsilon_8\varepsilon_{11}\}$, and the set of markings consistent with w is $\mathcal{C}(w) = \{[0\ 0\ 1\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 0]^T, [0\ 0\ 0\ 1\ 0\ 0\ 0\ 1\ 0\ 0\ 0]^T, [0\ 0\ 0\ 0\ 1\ 0\ 0\ 1\ 0\ 0\ 0]^T, [0\ 1\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 0]^T, [0\ 0\ 0\ 0\ 0\ 1\ 0\ 1\ 0\ 0\ 0]^T\}$.

If we consider $w = acd$ the set of firing sequences that are consistent with w is $\mathcal{S}(w) = \{t_1t_5t_6, t_1t_5\varepsilon_{12}\varepsilon_{13}t_7\}$, and the set of markings consistent with w is $\mathcal{C}(w) = \{[0\ 1\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 0]^T\}$. Thus two different firing sequences may have fired (the second one also involving silent transitions), but they both lead to the same marking. ■

4 MINIMAL EXPLANATIONS AND MINIMAL E-VECTORS

In this section we present the notions of minimal explanations and minimal e-vectors for labeled PNs. First we introduce notions of explanations for unlabeled PNs, secondly we define when an explanation is minimal and finally we extend these concepts to labeled PN.

Definition 4.1 (Cabasino et al., 2008). Given a marking M and an observable transition $t \in T_o$, we define

$$\Sigma(M,t) = \{\sigma \in T_u^* \mid M[\sigma\rangle M',\ M' \geq Pre(\cdot,t)\}$$

the set of *explanations* of t at M, and

$$Y(M,t) = \pi(\Sigma(M,t))$$

the *e-vectors* (or *explanation vectors*), i.e., firing vectors associated to the explanations. ■

Thus $\Sigma(M,t)$ is the set of unobservable sequences whose firing at M enables t. Among the above sequences we want to select those whose firing vector is minimal. The firing vector of these sequences are called *minimal e-vectors*.

Definition 4.2 (Cabasino et al., 2008). Given a marking M and a transition $t \in T_o$, we define

$$\Sigma_{\min}(M,t) = \{\sigma \in \Sigma(M,t) \mid \nexists\, \sigma' \in \Sigma(M,t) : \pi(\sigma') \lneqq \pi(\sigma)\}$$

the set of *minimal explanations* of t at M, and we define

$$Y_{\min}(M,t) = \pi(\Sigma_{\min}(M,t))$$

the corresponding set of *minimal e-vectors*. ■

In (Corona et al., 2004) we proved that, if the unobservable subnet is acyclic and backward conflict-free, then $|Y_{\min}(M,t)| = 1$.

Different approaches can be used to compute $Y_{\min}(M,t)$, e.g., (Boel and Jiroveanu, 2004; Jiroveanu and Boel, 2004). In (Cabasino et al., 2008) we suggested an approach that terminates finding all vectors in $Y_{\min}(M,t)$ if applied to nets whose unobservable subnet is acyclic. It simply requires algebraic manipulations, and is inspired by the procedure proposed in (Martinez and Silva, 1982) for the computation of minimal P-invariants. For the sake of brevity, this algorithm is not reported here.

In the case of labeled PNs what we observe are symbols in L. Thus, it is useful to compute the following sets.

Definition 4.3 (Cabasino et al., 2009). Given a marking M and an observation $l \in L$, we define the set of *minimal explanations of l at M* as

$$\hat{\Sigma}_{\min}(M,l) = \cup_{t \in T_l} \cup_{\sigma \in \Sigma_{\min}(M,t)} (t,\sigma),$$

i.e., the set of pairs (transition labeled l; corresponding minimal explanation), and we define the set of *minimal e-vectors of l at M* as

$$\hat{Y}_{\min}(M,l) = \cup_{t \in T_l} \cup_{e \in Y_{\min}(M,t)} (t,e),$$

i.e., the set of pairs (transition labeled l; corresponding minimal e-vector). ■

Thus, $\hat{\Sigma}_{\min}(M,l)$ is the set of pairs whose first element is the transition labeled l and whose second element is the corresponding minimal explanation $\sigma \in \Sigma_{\min}(M,t)$, namely the corresponding sequence of unobservable transitions whose firing at M enables l and whose firing vector is minimal. Moreover, $\hat{Y}_{\min}(M,l)$ is the set of pairs whose first element is the transition labeled l and whose second element is the firing vector $e \in Y_{\min}(M,t)$ corresponding to the second element in $\hat{\Sigma}_{\min}(M,l)$.

Obviously, $\hat{\Sigma}_{\min}(M,l)$ and $\hat{Y}_{\min}(M,l)$ are a generalization of the sets of minimal explanations and minimal e-vectors introduced for unlabeled PNs with unobservable transitions. Moreover, in the above sets $\hat{\Sigma}_{\min}(M,l)$ and $\hat{Y}_{\min}(M,l)$ different sequences σ and different e-vectors e, respectively, are associated in general to the same $t \in T_l$.

5 BASIS MARKINGS AND J-VECTORS

In this section we introduce the definitions of basis markings and justifications that are the crucial notions of our diagnosis approach.

In particular, given a sequence of observed events $w \in L^*$, a basis marking M_b is a marking reached from M_0 with the firing of the observed word w and of all unobservable transitions whose firing is necessary to enable w. Note that, in general several sequences $\sigma_o \in T_o^*$ may correspond to the same w, i.e., there are several sequences of observable transitions such that $\mathcal{L}(\sigma_o) = w$ that may have actually fired. Moreover, in general, to any of such sequences σ_o a different sequence of unobservable transitions interleaved with it is necessary to make it firable at the initial marking. Thus we need to introduce the following definition of pairs (sequence of transitions in T_o labeled w; corresponding *justification*).

Definition 5.1 (Cabasino et al., 2009). Let $\langle N, M_0 \rangle$ be a net system with labeling function $\mathcal{L} : T \to L \cup \{\varepsilon\}$, where $N = (P,T,Pre,Post)$ and $T = T_o \cup T_u$. Let $w \in L^*$ be a given observation. We define

$$\begin{aligned}\hat{\mathcal{J}}(w) \;=\{\, &(\sigma_o,\sigma_u),\ \sigma_o \in T_o^*,\ \mathcal{L}(\sigma_o) = w,\ \sigma_u \in T_u^* \mid \\ &[\exists \sigma \in \mathcal{S}(w) : \sigma_o = P_o(\sigma),\ \sigma_u = P_u(\sigma)] \wedge \\ &[\nexists \sigma' \in \mathcal{S}(w) : \sigma_o = P_o(\sigma'),\ \sigma'_u = P_u(\sigma') \wedge \\ &\qquad \pi(\sigma'_u) \lneqq \pi(\sigma_u)]\}\end{aligned}$$

the set of pairs (sequence $\sigma_o \in T_o^*$ with $\mathcal{L}(\sigma_o) = w$; corresponding *justification* of w). Moreover, we define

$$\begin{aligned}\hat{Y}_{\min}(M_0,w) = \{&(\sigma_o,y),\ \sigma_o \in T_o^*, \mathcal{L}(\sigma_o) = w, y \in \mathbb{N}^{n_u} \mid \\ &\exists (\sigma_o,\sigma_u) \in \hat{\mathcal{J}}(w) : \pi(\sigma_u) = y\}\end{aligned}$$

the set of pairs (sequence $\sigma_o \in T_o^*$ with $\mathcal{L}(\sigma_o) = w$; corresponding *j-vector*). ■

In simple words, $\hat{\mathcal{J}}(w)$ is the set of pairs whose first element is the sequence $\sigma_o \in T_o^*$ labeled w and whose second element is the corresponding sequence of unobservable transitions interleaved with σ_o whose firing enables σ_o and whose firing vector is minimal.

The firing vectors of these sequences are called *j-vectors*.

Definition 5.2 (Cabasino et al., 2009). Let $\langle N, M_0 \rangle$ be a net system with labeling function $\mathcal{L} : T \to L \cup \{\varepsilon\}$, where $N = (P, T, Pre, Post)$ and $T = T_o \cup T_u$. Let w be a given observation and $(\sigma_o, \sigma_u) \in \hat{\mathcal{J}}(w)$ be a generic pair (sequence of observable transitions labeled w; corresponding minimal justification). The marking

$$M_b = M_0 + C_u \cdot y + C_o \cdot y', \qquad y = \pi(\sigma_u),\ y' = \pi(\sigma_o),$$

i.e., the marking reached firing σ_o interleaved with the minimal justification σ_u, is called *basis marking* and y is called its *j-vector* (or *justification-vector*). ■

Obviously, because in general more than one justification exists for a word w (the set $\hat{\mathcal{J}}(w)$ is generally not a singleton), the basis marking may be not unique as well.

Definition 5.3 (Cabasino et al., 2009). Let $\langle N, M_0 \rangle$ be a net system with labeling function $\mathcal{L} : T \to L \cup \{\varepsilon\}$, where $N = (P, T, Pre, Post)$ and $T = T_o \cup T_u$. Let $w \in L^*$ be an observed word. We define

$$\begin{aligned} \mathcal{M}(w) = \{(M, y) \mid\ & (\exists \sigma \in \mathcal{S}(w) : M_0[\sigma\rangle M) \wedge \\ & (\exists (\sigma_o, \sigma_u) \in \hat{\mathcal{J}}(w) : \sigma_o = P_o(\sigma), \\ & \sigma_u = P_u(\sigma),\ y = \pi(\sigma_u))\} \end{aligned}$$

the set of pairs (basis marking; relative j-vector) that are *consistent* with $w \in L^*$. ■

Note that the set $\mathcal{M}(w)$ does not keep into account the sequences of observable transitions that may have actually fired. It only keeps track of the basis markings that can be reached and of the firing vectors relative to sequences of unobservable transitions that have fired to reach them. Indeed, this is the information really significant when performing diagnosis. The notion of $\mathcal{M}(w)$ is fundamental to provide a recursive way to compute the set of minimal explanations.

Proposition 5.4 (Cabasino et al., 2009). Given a net system $\langle N, M_0 \rangle$ with labeling function $\mathcal{L} : T \to L \cup \{\varepsilon\}$, where $N = (P, T, Pre, Post)$ and $T = T_o \cup T_u$. Assume that the unobservable subnet is acyclic. Let $w = w'l$ be a given observation.

The set $\hat{Y}_{\min}(M_0, wl)$ is defined as:

$$\begin{aligned} \hat{Y}_{\min}(M_0, wl) = \ & \{(\sigma_o, y) \mid \sigma_o = \sigma_0' t \wedge y = y' + e : \\ & (\sigma_o', y') \in \hat{Y}_{\min}(M_0, w), \\ & (t, e) \in \hat{Y}_{\min}(M_b', l) \text{ and } \mathcal{L}(t) = l\}, \end{aligned}$$

where $M_b' = M_0 + C_u \cdot y' + C_o \cdot \sigma_o'$.

Example 5.5. Let us consider the PN in Figure 1 previously introduced in Example 3.2.

Let us assume $w = acd$. The set of justifications is $\hat{\mathcal{J}}(w) = \{(t_1 t_5 t_6, \varepsilon), (t_1 t_5 t_7, \varepsilon_{12}\varepsilon_{13})\}$ and the set of j-vectors is $\hat{Y}_{min}(M_0, w) = \{(t_1 t_5 t_6, \vec{0}), (t_1 t_5 t_7, [0\ 0\ 0\ 0\ 1\ 1]^T)\}$. The above j-vectors lead to the same basis marking $M_b = [0\ 1\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 0]^T$ thus $\mathcal{M}(w) = \{(M_b, \vec{0}), (M_b, [0\ 0\ 0\ 0\ 1\ 1]^T)\}$.

Now, let us consider $w = ab$. In this case $\hat{\mathcal{J}}(w) = \{(t_1 t_2, \varepsilon)\}$, $\hat{Y}_{min}(M_0, w) = \{(t_1 t_2, \vec{0})\}$ and the basis marking is the same as in the previous case, namely $M_b = [0\ 1\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 0]^T$, thus $\mathcal{M}(w) = \{(M_b, \vec{0})\}$. ■

Under the assumption of acyclicity of the unobservable subnet, the set $\mathcal{M}(w)$ can be easily constructed as follows.

Algorithm 5.6 (Computation of the basis markings and j-vectors).

```
1. Let $w = \varepsilon$.
2. Let $\mathcal{M}(w) = \{(M_0, \vec{0})\}$.
3. Wait until a new label $l$ is observed.
4. Let $w' = w$ and $w = w'l$.
5. Let $\mathcal{M}(w) = \emptyset$.
6. For all $M'$ such that $(M', y') \in \mathcal{M}(w')$ , do
  6.1. for all $t \in T_l$, do
    6.1.1. for all $e \in Y_{\min}(M', t)$, do
    6.1.1.1. let $M = M' + C_u \cdot e + C(\cdot, t)$,
    6.1.1.2. for all $y'$ such that $(M', y') \in \mathcal{M}(w')$, do
        6.1.2.1. let $y = y' + e$,
        6.1.2.2. let $\mathcal{M}(w) = \mathcal{M}(w) \cup \{(M, y)\}$.
7. Goto step 3.
```

■

In simple words, the above algorithm can be explained as follows. We assume that a certain word w (that is equal to the empty string at the initial step) has been observed. Then, a new observable t fires and we observe its label $\mathcal{L}(t)$ (e.g., l). We consider all basis markings at the observation w' before the firing of t, and we select among them those that may have allowed the firing of at least one transition $t \in T_l$, also taking into account that this may have required the firing of appropriate sequences of unobservable transitions. In particular, we focus on the minimal explanations, and thus on the corresponding minimal e-vectors (step 6.1.1). Finally, we update the set $\mathcal{M}(w)$ including all pairs of new basis markings and j-vectors, taking into account that for each basis marking at w' it may correspond more than one j-vector.

Let us now recall the following result.

Definition 5.7 (Cabasino et al., 2008). Let $\langle N, M_0 \rangle$ be a net system where $N = (P, T, Pre, Post)$ and $T = T_o \cup T_u$. Assume that the unobservable subnet is acyclic. Let $w \in T_o^*$ be an observed word. We denote

$$\mathcal{M}_{basis}(w) = \{M \in \mathbb{N}^m \mid \exists y \in \mathbb{N}^{n_u} \text{ and } (M, y) \in \mathcal{M}(w)\}$$

the set of basis markings at w. Moreover, we denote as

$$\mathcal{M}_{basis} = \bigcup_{w \in T_o^*} \mathcal{M}_{basis}(w)$$

the set of all basis markings for any observation w. ■

Note that if the net system is bounded then the set $\mathcal{M}_{basis}$ is *finite* being the set of basis markings a subset of the reachability set.

Theorem 5.8 (Cabasino et al., 2008). Let us consider a net system $\langle N, M_0 \rangle$ whose unobservable subnet is acyclic. For any $w \in L^*$ it holds that

$$\mathcal{C}(w) = \{M \in \mathbb{N}^m \mid \quad M = M_b + C_u \cdot y \ : \ y \geq \vec{0} \ \text{ and } \ M_b \in \mathcal{M}_{basis}(w)\}.$$

6 DIAGNOSIS USING PETRI NETS

Assume that the set of unobservable transitions is partitioned into two subsets, namely $T_u = T_f \cup T_{reg}$ where T_f includes all fault transitions (modeling anomalous or fault behavior), while T_{reg} includes all transitions relative to unobservable but regular events. The set T_f is further partitioned into r different subsets T_f^i, where $i = 1, \ldots, r$, that model the different fault classes.

The following definition introduces the notion of *diagnoser*.

Definition 6.1 (Cabasino et al., 2009). A *diagnoser* is a function $\Delta : L^* \times \{T_f^1, T_f^2, \ldots, T_f^r\} \to \{0, 1, 2, 3\}$ that associates to each observation $w \in L^*$ and to each fault class T_f^i, $i = 1, \ldots, r$, a *diagnosis state*.

- $\Delta(w, T_f^i) = 0$ if for all $\sigma \in \mathcal{S}(w)$ and for all $t_f \in T_f^i$ it holds $t_f \notin \sigma$.
  In such a case the ith fault cannot have occurred, because none of the firing sequences consistent with the observation contains fault transitions of class i.
- $\Delta(w, T_f^i) = 1$ if:
  (i) there exist $\sigma \in \mathcal{S}(w)$ and $t_f \in T_f^i$ such that $t_f \in \sigma$ but
  (ii) for all $(\sigma_o, \sigma_u) \in \hat{\mathcal{J}}(w)$ and for all $t_f \in T_f^i$ it holds that $t_f \notin \sigma_u$.
  In such a case a fault transition of class i may have occurred but is not contained in any justification of w.
- $\Delta(w, T_f^i) = 2$ if there exist $(\sigma_o, \sigma_u), (\sigma_o', \sigma_u') \in \hat{\mathcal{J}}(w)$ such that
  (i) there exists $t_f \in T_f^i$ such that $t_f \in \sigma_u$;
  (ii) for all $t_f \in T_f^i$, $t_f \notin \sigma_u'$.
  In such a case a fault transition of class i is contained in one (but not in all) justification of w.
- $\Delta(w, T_f^i) = 3$ if for all $\sigma \in \mathcal{S}(w)$ there exists $t_f \in T_f^i$ such that $t_f \in \sigma$.
  In such a case the ith fault must have occurred, because all firable sequences consistent with the observation contain at least one fault in T_f^i. ■

Example 6.2. Let us consider the PN in Figure 1 previously introduced in Example 3.2. Let $T_f = \{\varepsilon_{11}, \varepsilon_{12}\}$. Assume that the two fault transitions belong to different fault classes, i.e., $T_f^1 = \{\varepsilon_{11}\}$ and $T_f^2 = \{\varepsilon_{12}\}$.

Let us observe $w = a$. Then $\Delta(w, T_f^1) = \Delta(w, T_f^2) = 0$, being $\hat{\mathcal{J}}(w) = \{(t_1, \varepsilon)\}$ and $\mathcal{S}(w) = \{t_1\}$. In words no fault of both fault classes can have occurred.

Let us observe $w = ab$. Then $\Delta(w, T_f^1) = 1$ and $\Delta(w, T_f^2) = 0$, being $\hat{\mathcal{J}}(w) = \{(t_1 t_2, \varepsilon)\}$ and $\mathcal{S}(w) = \{t_1 t_2, t_1 t_2 \varepsilon_8, t_1 t_2 \varepsilon_8 \varepsilon_9, t_1 t_2 \varepsilon_8 \varepsilon_9 \varepsilon_{10}, t_1 t_2 \varepsilon_8 \varepsilon_{11}\}$. This means that a fault of the second fault class may have occurred (e.g. $t_1 t_2 \varepsilon_8 \varepsilon_{11}$) but it is not contained in any justification of ab, while no fault of the first fault class can have occurred.

Now, let us consider $w = abb$. In this case $\Delta(w, T_f^1) = 2$ and $\Delta(w, T_f^2) = 0$, being $\hat{\mathcal{J}}(w) = \{(t_1 t_2 t_2, \varepsilon_8 \varepsilon_9 \varepsilon_{10}), (t_1 t_2 t_3, \varepsilon_8 \varepsilon_{11})\}$ and $\mathcal{S}(w) = \{t_1 t_2 \varepsilon_8 \varepsilon_9 \varepsilon_{10} t_2, t_1 t_2 \varepsilon_8 \varepsilon_9 \varepsilon_{10} t_2 \varepsilon_8, t_1 t_2 \varepsilon_8 \varepsilon_9 \varepsilon_{10} t_2 \varepsilon_8 \varepsilon_9, t_1 t_2 \varepsilon_8 \varepsilon_9 \varepsilon_{10} t_2 \varepsilon_8 \varepsilon_9 \varepsilon_{10}, t_1 t_2 \varepsilon_8 \varepsilon_9 \varepsilon_{10} t_2 \varepsilon_8 \varepsilon_{11}\}$. This means that no fault of the first fault class can have occurred, while a fault of the second fault class may have occurred since one justification does not contain ε_{11} and one justification contains it.

Finally, let us consider $w = abbccc$. In this case $\Delta(w, T_f^1) = 1$ and $\Delta(w, T_f^2) = 3$. In fact since $\hat{\mathcal{J}}(w) = \{(t_1 t_2 t_3 t_5 t_4 t_4, \varepsilon_8 \varepsilon_{11}), (t_1 t_2 t_3 t_4 t_5 t_4, \varepsilon_8 \varepsilon_{11}), (t_1 t_2 t_3 t_4 t_4 t_5, \varepsilon_8 \varepsilon_{11}), (t_1 t_2 t_3 t_4 t_4 t_4, \varepsilon_8 \varepsilon_{11})\}$ a fault of the first fault class must have occurred, while a fault of the second fault class may have occurred (e.g. $t_1 t_2 \varepsilon_8 \varepsilon_{11} t_3 t_4 t_4 t_5 \varepsilon_{12}$) but it is not contained in any justification of w. ■

The following proposition presents how the diagnosis states can be characterized analyzing basis markings and justifications.

Proposition 6.3 (Cabasino et al., 2009). Consider an observed word $w \in L^*$.

- $\Delta(w, T_f^i) \in \{0, 1\}$ iff for all $(M, y) \in \mathcal{M}(w)$ and for all $t_f \in T_f^i$ it holds $y(t_f) = 0$.
- $\Delta(w, T_f^i) = 2$ iff there exist $(M, y) \in \mathcal{M}(w)$ and $(M', y') \in \mathcal{M}(w)$ such that:
  (i) there exists $t_f \in T_f^i$ such that $y(t_f) > 0$,
  (ii) for all $t_f \in T_f^i$, $y'(t_f) = 0$.

- $\Delta(w,T_f^i)=3$ iff for all $(M,y)\in\mathcal{M}(w)$ there exists $t_f\in T_f^i$ such that $y(t_f)>0$.

The following proposition shows how to distinguish between diagnosis states 0 and 1.

Proposition 6.4 (Cabasino et al., 2009). For a PN whose unobservable subnet is acyclic, let $w\in L^*$ be an observed word such that for all $(M,y)\in\mathcal{M}(w)$ it holds $y(t_f)=0\ \forall\ t_f\in T_f^i$. Let us consider the constraint set

$$\mathcal{T}(M) \quad = \quad \left\{\begin{array}{l} M+C_u\cdot z\geq\vec{0},\\ \displaystyle\sum_{t_f\in T_f^i} z(t_f)>0,\\ z\in\mathbb{N}^{n_u}.\end{array}\right. \tag{1}$$

- $\Delta(w,T_f^i)=0$ if $\forall\ (M,y)\in\mathcal{M}(w)$ the constraint set (1) is not feasible.
- $\Delta(w,T_f^i)=1$ if $\exists\ (M,y)\in\mathcal{M}(w)$ such that the constraint set (1) is feasible.

On the basis of the above two results, if the unobservable subnet is acyclic, diagnosis may be carried out by simply looking at the set $\mathcal{M}(w)$ for any observed word w and, should the diagnosis state be either 0 or 1, by additionally evaluating whether the corresponding integer constraint set (1) admits a solution.

Example 6.5. Let us consider the PN in Figure 1 where $T_f^1=\{\varepsilon_{11}\}$ and $T_f^2=\{\varepsilon_{12}\}$.

Let $w=ab$. In this case $\mathcal{M}(w)=\{(M_b^1,\vec{0})\}$, where $M_b^1=[0\,1\,0\,0\,0\,0\,0\,1\,0\,0\,0]^T$. Being $\mathcal{T}(M_b^1)$ feasible only for the fault class T_f^1 it holds $\Delta(w,T_f^1)=1$ and $\Delta(w,T_f^2)=0$.

Let $w=abb$. It is $\mathcal{M}(w)=\{(M_b^1,[1\ 1\ 1\ 0\ 0\ 0]^T)$, $(M_b^2,[1\ \ 0\ \ 0\ \ 1\ \ 0\ \ 0]^T)\}$, where $M_b^2=[0\ 0\ 0\ 0\ 0\ 0\ 1\ 1\ 0\ 0\ 0]^T$. It is $\Delta(w,T_f^1)=2$ and $\Delta(w,T_f^2)=0$ being both $\mathcal{T}(M_b^1)$ and $\mathcal{T}(M_b^2)$ not feasible.

Let $w=abbccc$. In this case $\mathcal{M}(w)=\{(M_b^3,[1\ 1\ 1\ 0\ 0\ 0]^T),(M_b^4,[1\ 1\ 1\ 0\ 0\ 0]^T)\}$, where $M_b^3=[0\ 0\ 0\ 0\ 0\ 0\ 1\ 1\ 0\ 0\ 0]^T$ and $M_b^4=[0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 1\ 0\ 0]^T$. It is $\Delta(w,T_f^1)=3$ and being $\mathcal{T}(M_b^4)$ feasible for the second fault class T_f^2 it holds $\Delta(w,T_f^2)=1$. ■

7 BASIS REACHABILITY GRAPH

Diagnosis approach described in the previous section can be applied both to bounded and unbounded PNs. The proposed approach is an on-line approach that for each new observed event updates the diagnosis state for each fault class computing the set of basis markings and j-vectors. Moreover if for a given fault class is necessary to distinguish between diagnosis states 0 and 1, it is also necessary to solve for each basis marking M_b the constraint set $\mathcal{T}(M_b)$.

In this section we show that if the considered net system is bounded, the most burdensome part of the procedure can be moved off-line defining a graph called *Basis Reachability Graph* (BRG).

Definition 7.1. The BRG is a deterministic graph that has as many nodes as the number of possible basis markings.

To each node is associated a different basis marking M and a row vector with as many entries as the number of fault classes. The entries of this vector may only take binary values: 1 if $\mathcal{T}(M)$ is feasible, 0 otherwise.

Arcs are labeled with observable events in L and e-vectors. More precisely, an arc exists from a node containing the basis marking M to a node containing the basis marking M' if and only if there exists a transition t for which an explanation exists at M and the firing of t and one of its minimal explanations leads to M'. The arc going from M to M' is labeled $(\mathcal{L}(t),e)$, where $e\in Y_{\min}(M,t)$ and $M'=M+C_u\cdot e+C(\cdot,t)$. ■

Note that the number of nodes of the BRG is always finite being the set of basis markings a subset of the set of reachable markings, that is finite being the net bounded. Moreover, the row vector of binary values associated to the nodes of the BRG allows us to distinguish between the diagnosis state 1 or 0.

The main steps for the computation of the BRG in the case of labeled PNs are summarized in the following algorithm.

Algorithm 7.2 (Computation of the BRG).

```
1. Label the initial node $(M_0,x_0)$ where $\forall i=1,\ldots,r$,
     $x_0(T_f^i)=\begin{cases}1 & \text{if } \mathcal{T}(M_0) \text{ is feasible,}\\ 0 & \text{otherwise.}\end{cases}$
   Assign no tag to it.
2. While nodes with no tag exist
   select a node with no tag and do
   2.1. let $M$ be the marking in the node $(M,x)$,
   2.2. for all $l\in L$
     2.2.1. for all $t: L(t)=l\wedge Y_{\min}(M,t)\neq\emptyset$, do
            • for all $e\in Y_{\min}(M,t)$, do
              • let $M'=M+C_u\cdot e+C(\cdot,t)$,
              • if $\nexists$ a node $(M,x)$ with $M=M'$, do
                • add a new node to the graph containing
                  $(M',x')$ where $\forall i=1,\ldots,r$,
                  $x'(T_f^i)=\begin{cases}1 & \text{if } \mathcal{T}(M') \text{ is feasible,}\\ 0 & \text{otherwise.}\end{cases}$
                  and arc $(l,e)$ from $(M,x)$ to $(M',x')$
              • else
                • add arc $(l,e)$ from $(M,x)$ to $(M',x')$
            if it does not exist yet
    2.3. tag the node ”old”.
3. Remove all tags.
```

■

The algorithm constructs the BRG starting from the initial node to which it corresponds the initial marking and a binary vector defining which classes of faults may occur at M_0. Now, we consider all the labels $l \in L$ such that there exists a transition t with $L(t) = l$ for which a minimal explanation at M_0 exists. For any of these transitions we compute the marking resulting from firing t at $M_0 + C_u \cdot e$, for any $e \in Y_{\min}(M_0, t)$. If a pair (marking, binary vector) not contained in the previous nodes is obtained, a new node is added to the graph. The arc going from the initial node to the new node is labeled (l, e). The procedure is iterated until all basis markings have been considered. Note that, our approach always requires to enumerate a state space that is a strict subset of the reachability space. However, as in general for diagnosis approaches, the combinatory explosion cannot be avoided.

Example 7.3. Let us consider the PN in Figure 1, where $T_o = \{t_1, t_2, t_3, t_4, t_5, t_6, t_7\}$, $T_u = \{\varepsilon_8, \varepsilon_9, \varepsilon_{10}, \varepsilon_{11}, \varepsilon_{12}, \varepsilon_{13}\}$, $T_f^1 = \{\varepsilon_{11}\}$ and $T_f^2 = \{\varepsilon_{12}\}$. The labeling function is defined as follows: $\mathcal{L}(t_1) = a$, $\mathcal{L}(t_2) = \mathcal{L}(t_3) = b$, $\mathcal{L}(t_4) = \mathcal{L}(t_5) = c$, $\mathcal{L}(t_6) = \mathcal{L}(t_7) = d$.

The BRG is shown in Figure 2. The notation used in in this figure is detailed in Tables 1 and 2. Each node contains a different basis marking and a binary row vector of dimension two, being two the number of fault classes. As an example, the binary vector [0 0] is associated to M_0 because $\mathcal{T}(M_0)$ is not feasible for both fault classes. From node M_0 to node M_1 there is one arc labeled a and with the null vector as minimal explanation. The node containing the basis marking M_2 has binary vector [0 1], because $\mathcal{T}(M_2)$ is feasible only for T_f^2. Node $(M_2, [0\ 1])$ has two output arcs both labeled with d and both directed to node $(M_1, [0\ 0])$ with two different minimal explanations $\vec{0}$ and e_1, respectively, plus another output arc $(b, \vec{0})$ directed to node $(M_4, [1\ 1])$. ■

The following algorithm summarizes the main steps of the on-line diagnosis carried out by looking at the BRG.

Algorithm 7.4 (Diagnosis using the BRG).

```
1. Let $w = \varepsilon$.
2. Let $\mathcal{M}(w) = \{(M_0, \vec{0})\}$.
3. Wait until a new observable transition fires.
   Let $l$ be the observed event.
4. Let $w' = w$ and $w = w'l$.
5. Let $\mathcal{M}(w) = \emptyset$,            [Computation of $\mathcal{M}(w)$]
6. For all nodes containing $M' : (M', y') \in \mathcal{M}(w')$, do
  6.1. for all arcs exiting from the node with $M'$, do
    6.1.1. let $M$ be the marking of the output node
           and $e$ be the minimal e-vector on the edge
           from $M'$ to $M$,
    6.1.2. for all $y'$ such that $(M', y') \in \mathcal{M}(w')$, do
      6.1.2.1. let $y = y' + e$,
      6.1.2.2. let $\mathcal{M}(w) = \mathcal{M}(w) \cup \{(M, y)\}$,
7. for all $i = 1, \ldots, r$, do
                          [Computation of the diagnosis state]
  7.1. if $\forall\ (M, y) \in \mathcal{M}(w) \wedge \forall t_f \in T_f^i$ it is $y(t_f) = 0$, do
    7.1.1. if $\forall\ (M, y) \in \mathcal{M}(w)$ it holds $x(i) = 0$,
           where $x$ is the binary vector in node $M$, do
      7.1.1.1. let $\Delta(w, T_f^i) = 0$,
    7.1.2. else
      7.1.2.1. let $\Delta(w, T_f^i) = 1$,
  7.2. if $\exists\ (M, y) \in \mathcal{M}(w)$ and $(M', y') \in \mathcal{M}(w)$ s.t.:
       (i) $\exists t_f \in T_f^i$ such that $y(t_f) > 0$,
       (ii) $\forall t_f \in T_f^i, y'(t_f) = 0$, do
    7.2.1. let $\Delta(w, T_f^i) = 2$,
  7.3. if $\forall\ (M, y) \in \mathcal{M}(w)\ \exists t_f \in T_f^i : y(t_f) > 0$, do
    7.3.1. let $\Delta(w, T_f^i) = 3$.
8. Goto step 3.
```

■

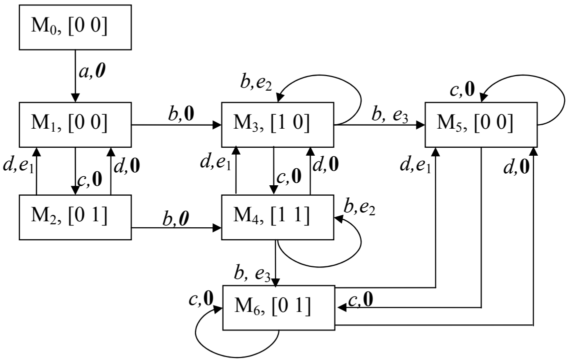


Figure 2: The BRG of the PN in Figure 1.

Steps 1 to 6 of Algorithm 7.4 enables us to compute the set $\mathcal{M}(w)$. When no event is observed, namely $w = \varepsilon$, then $\mathcal{M}(w) = \{(M_0, \vec{0})\}$. Now, assume that a label l is observed. We include in the set $\mathcal{M}(l)$ all couples (M, y) such that an arc labeled l exits from the initial node and ends in a node containing the basis marking M. The corresponding value of y is equal to the e-vector in the arc going from M_0 to M, being $\vec{0}$ the j-vector relative to M_0. In general, if w' is the actual observation, and a new event labeled l fires, we consider all couples $(M', y') \in \mathcal{M}(w')$ and all nodes that can be reached from M' with an arc labeled l. Let M be the basis marking of the generic resulting node. We include in $\mathcal{M}(w) = \mathcal{M}(w't)$ all couples (M, y), where for any M, y is equal to the sum of y' plus the e-vector labeling the arc from M' to M.

Step 7 of Algorithm 7.4 computes the diagnosis

Table 1: The markings of the BRG in Figure 2.

M_0	$[\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ ]^T$
M_1	$[\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 0\ ]^T$
M_2	$[\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ ]^T$
M_3	$[\ 0\ 0\ 1\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 0\ ]^T$
M_4	$[\ 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ ]^T$
M_5	$[\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 1\ 0\ 0\ 0\ ]^T$
M_6	$[\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 1\ 0\ 0\ ]^T$

Table 2: The e-vectors of the BRG in Figure 2.

	ε_8	ε_9	ε_{10}	ε_{11}	ε_{12}	ε_{13}
e_1	0	0	0	0	1	1
e_2	1	1	1	0	0	0
e_3	1	0	0	1	0	0

state. Let us consider the generic ith fault class. If $\forall (M,y) \in \mathcal{M}(w)$ and $\forall t_f \in T_f^i$ it holds $y(t_f) = 0$, we have to check the ith entry of all the binary row vectors associated to the basis markings M, such that $(M,y) \in \mathcal{M}(w)$. If these entries are all equal to 0, we set $\Delta(w, T_f^i) = 0$, otherwise we set $\Delta(w, T_f^i) = 1$. On the other hand, if there exists at least one pair $(M,y) \in \mathcal{M}(w)$ with $y(t_f) > 0$ for any $t_f \in T_f^i$, and there exists at least one pair $(M',y') \in \mathcal{M}(w)$ with $y(t_f) = 0$ for all $t_f \in T_f^i$, then $\Delta(w, T_f^i) = 2$. Finally, if for all pairs $(M,y) \in \mathcal{M}(w)$ $y(t_f) > 0$ for any $t_f \in T_f^i$, then $\Delta(w, T_f^i) = 3$.

The following example shows how to perform diagnosis on-line simply looking at the BRG.

Example 7.5. Let us consider the PN in Figure 1 and its BRG in Figure 2. Let $w = \varepsilon$. By looking at the BRG we establish that $\Delta(\varepsilon, T_f^1) = \Delta(\varepsilon, T_f^2) = 0$ being both entries of the row vector associated to M_0 equal to 0.

Now, let us consider $w = ab$. In such a case $\mathcal{M}(w) = \{(M_3, \vec{0})\}$. It holds $\Delta(ab, T_f^1) = 1$ and $\Delta(ab, T_f^2) = 0$ being the row vector in the node equal to [1 0].

Finally, for $w = abbc$ it holds $\Delta(abbc, T_f^1) = 2$ and $\Delta(abbc, T_f^2) = 1$. In fact $\mathcal{M}(w) = \{(M_4, y_1), (M_5, y_2)\}$, where $y_1 = e_2$, $y_2 = e_2 + e_3$, and the row vectors associated to M_4 and M_5 are respectively [1 1] and [0 0]. ■

8 MATLAB TOOLBOX

Our group at the University of Cagliari has developed a MATLAB toolbox for PNs.

In this section we illustrate how it can be used for the diagnosis of labeled PNs. In particular, we consider the function that given a bounded labeled PN builds the basis reachability graph.

The input of the MATLAB function BRG.m are:

- the structure of the net, i.e., the matrices *Pre* and *Post*;
- the initial marking M_0;
- a cell array F that has as many rows as the number of fault classes, that contains in each row the fault transitions that belong to the corresponding fault class;
- a cell array L that has as many rows as the cardinality of the considered alphabet, that contains in each row the observable transitions having the same label;
- a cell array E that contains in each row a string of characters, each one corresponding to a different label in the considered alphabet. Obviously, the cell array E is ordered according to L.

The output of the MATLAB function BRG.m is a cell array T that univocally identifies the resulting BRG. It has as many rows as the number of nodes of the BRG. A different row is associated to each node and contains the following information:

- an identifier number of the node;
- a matrix whose rows are equal to the transpose of the basis markings associated to the node;
- a matrix with as many rows as the number of basis markings associated to the node and as many columns as the number of fault classes: the jth element in the ith row (corresponding to M_b^i) is equal to $x_i(T_f^j)$ evaluated at M_b^i. Thus, $x_i(T_f^j) = 0$ is $\mathcal{T}(M_b^i)$ is not feasible with respect to T_j^f, 1 otherwise;

- the transitions enabled at node;
- the identifier number of the nodes that are reached firing an enabled transition and the corresponding j-vector.

9 NUMERICAL SIMULATIONS

Let us consider the Petri net in Figure 3 (Lai et al., 2008), where thick transitions represent observable event and thin transitions represent unobservable events. It models a family of manufacturing systems characterized by three parameters: n, m and k.

— n is the number of production lines.

— m is the number of units of the final product that can be simultaneously produced. Each unit of product is composed of n parts.

— k is the number of operations that each part must undergo in each line.

To obtain one unit of final product n orders are sent, one to each line; this is represented by observable event t_s. Each line will produce a part (all parts are identical) and put it in its final buffer. An assembly station will take one part from each buffer (observable event t_e) to produce the final product.

The part in line i ($i = 1, \dots, n$) undergoes a series of k operations, represented by unobservable events $\varepsilon_{i,1}, \varepsilon_{i,2}, \cdots, \varepsilon_{i,k}$.

After this series of operations two events are possible: either the part is regularly put in the final buffer of the line, or a fault may occur.

— Putting the part in the final buffer of line 1 corresponds to unobservable event $\varepsilon_{1,k+1}$, while putting the part in the final buffer of line i ($i = 2, \dots, n$) corresponds to observable event $t_{i,k+1}$.

— There are $n-1$ faults, represented by unobservable events f_i ($i = 1, \dots, n-1$). Fault f_i moves a part from line i to line $i+1$. Note that on line i ($i = 1, \dots, n-1$) the fault may only occur when the part has finished processing and is ready to be put in its final buffer; the part goes to the same processing stage in line $i+1$.

In this section we present the results of the computation of the BRG for several numerical simulations. Results obtained for different values of n, k and m are summarized in Tables 3, 4 and 5.

Note that for the sake of simplicity we assumed that all faults belong to the same class.

In these tables we also detail the cardinality of the reachability set R. This is an extremely important parameter to appreciate the advantage of using basis markings. The value of $|R|$ has been computed using a function we developed in MATLAB. For completeness we also reported the time necessary to compute it.

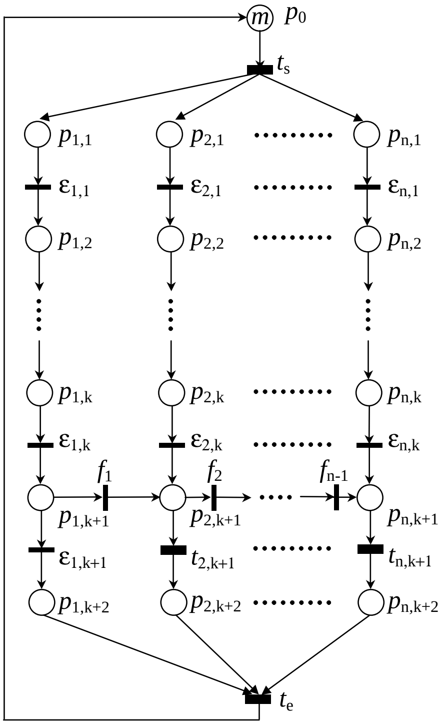


Figure 3: A manufacturing system.

Let us observe that some boxes of the above tables contain the non numerical values o:t: (out of time), that denotes that the corresponding value has not been computed within 6 hours.

All simulations have been run on a PC Athlon 64, 4000+ processor.

— Columns 1 and 2 show the values of n and k.

— Column 3 shows the number of nodes $|R|$ of the reachability graph.

— Column 4 shows the time t_R in seconds we spent to compute the reachability graph.

— Column 4 shows the number of nodes $|BRG|$ of the BRG.

— Column 5 shows the time t_{BRG} in seconds we spent to compute the BRG using the function BRG.m.

Tables 3, 4 and 5 show that the time spent to compute the reachability graph highly increases with the dimension of the net, namely with n and k, and with the number of products m.

On the contrary, the time spent to compute the BRG is always reasonable even for high values of n, k and m.

Tables 3, 4 and 5 also show that the number of nodes of the BRG only depends on n and m, while it is invariant with respect to k. On the other hand, $|R|$ also highly increases with k.

Table 3: Numerical results in the case of $m = 1$.

n	k	$\|R\|$	t_R [sec]	$\|BRG\|$	t_{BRG} [sec]
2	1	15	0.031	5	0.062
2	2	24	0.031	5	0.062
2	3	35	0.047	5	0.062
2	4	48	0.062	5	0.07
2	5	63	0.078	5	0.07
2	6	80	0.094	5	0.07
3	1	80	0.094	17	0.101
3	2	159	0.25	17	0.101
3	3	274	0.672	17	0.109
3	4	431	1.72	17	0.117
3	5	636	3.938	17	0.125
3	6	895	8.328	17	0.132
4	1	495	2.375	69	0.375
4	2	1200	16.969	69	0.43
4	3	2415	77.828	69	0.477
4	4	4320	272.53	69	0.531
4	5	7119	824.69	69	0.594
4	6	11040	2122.4	69	0.664
5	1	3295	155.81	305	4.345
5	2	9691	1615.7	305	4.765
5	3	22707	10288	305	5.25
5	4	*o.t.*	*o.t.*	305	5.75
5	5	*o.t.*	*o.t.*	305	6.897
5	6	*o.t.*	*o.t.*	305	7.894

Table 4: Numerical results in the case of $m = 2$.

n	k	$\|R\|$	t_R [sec]	$\|BRG\|$	t_{BRG} [sec]
2	1	96	0.11	17	0.086
2	2	237	0.469	17	0.094
2	3	496	2.078	17	0.1
3	1	1484	24.204	140	0.78
3	2	5949	486.39	140	0.844
3	3	18311	5320.9	140	0.906
4	1	28203	14006	1433	73.5
4	2	*o.t.*	*o.t.*	1433	76.5
4	3	*o.t.*	*o.t.*	1433	76.5

For the considered Petri net, on the basis of the above simulations, we can conclude that the diagnosis approach here presented is suitable from a computational point of view. In fact, thanks to the basis markings the reachability space can be described in a compact manner.

Table 5: Numerical results in the case of $m = 3$.

n	k	$\|R\|$	t_R [sec]	$\|BRG\|$	t_{BRG} [sec]
2	1	377	1.203	39	0.145
2	2	1293	17.203	39	0.145
3	1	12048	2113.9	553	8.219
3	2	*o.t.*	*o.t.*	553	9.016
4	1	*o.t.*	*o.t.*	9835	4095.06
4	2	*o.t.*	*o.t.*	9835	4095.06

10 CONCLUSIONS AND FUTURE WORK

This paper presents a diagnosis approach for labeled PNs using basis markings. This enables us to avoid an exhaustive enumeration of the reachability set. This approach applies to all bounded and unbounded Petri net systems whose unobservable subnet is acyclic. However, if we consider bounded net systems the most burdensome part of the procedure may be moved off-line computing the Basis Reachability Graph. Finally, we have presented a tool for the diagnosis of labeled bounded PNs and we have shown the simulation results using as diagnosis benchmark a family of manufacturing systems.

We have also studied the problem of diagnosability of bounded and unbounded PNs giving for both cases necessary and sufficient conditions for diagnosability. These results are not reported here, but they have been already submitted to an international conference.

Our future work will be that of studying the diagnosis problem for distributed systems investigating the possibility of extending the approach here presented to this case.

ACKNOWLEDGEMENTS

We thank Marco Pocci, a Master student of Electronic Engineering at the University of Cagliari, for his help in the development of the MATLAB tool for the construction of the BRG for labeled PNs.

REFERENCES

Basile, F., Chiacchio, P., and Tommasi, G. D. (2008). An efficient approach for online diagnosis of discrete event systems. *IEEE Trans. on Automatic Control*. in press.

Benveniste, A., Fabre, E., Haar, S., and Jard, C. (2003). Diagnosis of asynchronous discrete event systems: A net unfolding approach. *IEEE Trans. on Automatic Control*, 48(5):714–727.

Boel, R. and Jiroveanu, G. (2004). Distributed contextual diagnosis for very large systems. In *Proc. IFAC WODES'04: 7th Work. on Discrete Event Systems*, pages 343–348.

Boel, R. and van Schuppen, J. (2002). Decentralized failure diagnosis for discrete-event systems with costly communication between diagnosers. In *Proc. WODES'02: 6th Work. on Discrete Event Systems*, pages 175–181.

Cabasino, M., Giua, A., and Seatzu, C. (2008). Fault detection for discrete event systems using Petri nets with unobservable transitions. *Automatica*. Preliminary accepted.

Cabasino, M., Giua, A., and Seatzu, C. (2009). Diagnosis of discrete event systems using labeled Petri nets. In *Proc. 2nd IFAC Workshop on Dependable Control of Discrete Systems (Bari, Italy)*.

Chung, S. (2005). Diagnosing pn-based models with partial observable transitions. *International Journal of Computer Integrated Manufacturing*, 12 (2):158–169.

Corona, D., Giua, A., and Seatzu, C. (2004). Marking estimation of Petri nets with silent transitions. In *Proc. IEEE 43rd Int. Conf. on Decision and Control (Atlantis, The Bahamas)*.

Debouk, R., Lafortune, S., and Teneketzis, D. (2000). Coordinated decentralized protocols for failure diagnosis of discrete-event systems. *Discrete Events Dynamical Systems*, 10(1):33–86.

Dotoli, M., Fanti, M., and Mangini, A. (2008). Fault detection of discrete event systems using Petri nets and integer linear programming. In *Proc. of 17th IFAC World Congress*, Seoul, Korea.

Genc, S. and Lafortune, S. (2007). Distributed diagnosis of place-bordered Petri nets. *IEEE Trans. on Automation Science and Engineering*, 4(2):206–219.

Ghazel, M., Togueni, A., and Bigang, M. (2005). A monitoring approach for discrete events systems based on a time Petri net model. In *Proc. of 16th IFAC World Congress*, Prague, Czech Republic.

Giua, A. and Seatzu, C. (2005). Fault detection for discrete event systems using Petri nets with unobservable transitions. In *Proc. 44th IEEE Conf. on Decision and Control*, pages 6323–6328.

Hadjicostis, C. and Veghese, G. (1999). Monitoring discrete event systems using Petri net embeddings. *Lecture Notes in Computer Science*, 1639:188–207.

Jiroveanu, G. and Boel, R. (2004). Contextual analysis of Petri nets for distributed applications. In *16th Int. Symp. on Mathematical Theory of Networks and Systems (Leuven, Belgium)*.

Lai, S., Nessi, D., Cabasino, M., Giua, A., and Seatzu, C. (2008). A comparison between two diagnostic tools based on automata and Petri nets. In *Proc. IFAC WODES'08: 9th Work. on Discrete Event Systems*, pages 144–149.

Lefebvre, D. and Delherm, C. (2007). Diagnosis of DES with Petri net models. *IEEE Trans. on Automation Science and Engineering*, 4(1):114–118.

Lin, F. (1994). Diagnosability of discrete event systems and its applications. *Discrete Event Dynamic Systems*, 4(2):197–212.

Lin, F., Markee, J., , and Rado, B. (1993). Design and test of mixed signal circuits: a discrete event approach. In *Proc. 32rd IEEE Conf. on Decision and Control*, pages 246–251.

Martinez, J. and Silva, M. (1982). A simple and fast algorithm to obtain all invariants of a generalized Petri net. In *Informatik-Fachberichte 52: Application and Theory of Petri Nets.*, pages 301–310. Springer-Verlag.

Murata, T. (1989). Petri nets: properties, analysis and applications. *Proceedings of the IEEE*, 77(4):541–580.

Prock, J. (1991). A new tecnique for fault detection using Petri nets. *Automatica*, 27(2):239–245.

Ruiz-Beltràn, A. R.-T. E., Rivera-Rangel, I., and Lopez-Mellado, E. (2007). Online fault diagnosis of discrete event systems. A Petri net-based approach. *IEEE Trans. on Automation Science and Engineering*, 4(1):31–39.

Sampath, M., Lafortune, S., and Teneketzis, D. (1998). Active diagnosis of discrete-event systems. *IEEE Trans. on Automatic Control*, 43(7):908–929.

Sampath, M., Sengupta, R., Lafortune, S., Sinnamohideen, K., and Teneketzis, D. (1995). Diagnosability of discrete-event systems. *IEEE Trans. on Automatic Control*, 40 (9):1555–1575.

Sampath, M., Sengupta, R., Lafortune, S., Sinnamohideen, K., and Teneketzis, D. (1996). Failure diagnosis using discrete-event models. *IEEE Trans. Control Systems Technology*, 4(2):105–124.

Sreenivas, V. and Jafari, M. (1993). Fault detection and monitoring using time Petri nets. *IEEE Trans. Systems, Man and Cybernetics*, 23(4):1155–1162.

Ushio, T., Onishi, L., and Okuda, K. (1998). Fault detection based on Petri net models with faulty behaviors. In *Proc. SMC'98: IEEE Int. Conf. on Systems, Man, and Cybernetics (San Diego, CA, USA)*, pages 113–118.

Wen, Y. and Jeng, M. (2005). Diagnosability analysis based on T-invariants of Petri nets. In *Networking, Sensing and Control, 2005. Proceedings, 2005 IEEE.*, pages 371– 376.

Wen, Y., Li, C., and Jeng, M. (2005). A polynomial algorithm for checking diagnosability of Petri nets. In *Proc. SMC'05: IEEE Int. Conf. on Systems, Man, and Cybernetics*, pages 2542– 2547.

Wu, Y. and Hadjicostis, C. (2005). Algebraic approaches for fault identification in discrete-event systems. *IEEE Trans. Robotics and Automation*, 50(12):2048–2053.

Zad, S. H., Kwong, R., and Wonham, W. (2003). Fault diagnosis in discrete-event systems: framework and model reduction. *IEEE Trans. on Automatic Control*, 48(7):1199–1212.

BRIEF BIOGRAPHY

Alessandro Giua is professor of Automatic Control at the Department of Electrical and Electronic Engineering of the University of Cagliari, Italy. He received the Laurea degree in electric engineering from the University of Cagliari, Italy in 1988, and the M.S. and Ph.D. degrees in computer and systems engineering from Rensselaer Polytechnic Institute, Troy, New York, in 1990 and 1992.

His research interests include discrete event systems, hybrid systems, networked control systems, automated manufacturing, Petri nets, control of mechanical systems, failure diagnosis. He has co-authored two textbooks on Automatic Control (in Italian) and over 150 technical papers.

Dr. Giua is a member of the editorial board of the journals: Discrete Event Dynamic Systems: Theory and Applications; IEEE Trans. on Control Systems Technology; Nonlinear Analysis: Hybrid Systems. He has served in the program committee of over 60 international conferences.

He is chair for Chapter Activities of the Member Activities Board of the IEEE Control Systems Society and chair of the IFAC Technical Committee 1.3 on Discrete Event and Hybrid Systems.

MEETING THE WORLD CHALLENGES
From Philosophy to Information Technology to Applications

Peter Simon Sapaty
Institute of Mathematical Machines and Systems, National Academy of Sciences
Glushkova Ave 42, 03187 Kiev, Ukraine
sapaty@immsp.kiev.ua

Keywords: World crises, Atomism, holism, Gestalt theory, System integrity, Waves, Distributed scenario language, Networked interpretation, Unmanned systems, Task level, Behavioral level, Crisis management, Electronic warfare, Directed energy systems.

Abstract: We have been witnessing numerous world crises and disasters—from ecological to military to economic, with global world dynamics likely to be increasing this century further. The paper highlights known holistic and gestalt principles mainly used for a single brain, extending them to any distributed systems which may need high integrity and performance in reaction to unpredictable situations. A higher organizational layer is proposed enabling any distributed resources and systems to behave as an organism having global "consciousness" and pursuing global goals. This "over-operability" layer is established by implanting into key system points the same copy of a universal intelligent module, which can communicate with other such modules and interpret collectively global mission scenarios presented in a special Distributed Scenario Language. The scenarios can be injected from any module, and then self-replicate, self-modify, and self-spread throughout the system to be managed, tasking components, activating distributed resources, and establishing runtime infrastructures supporting system's integrity. Numerous existing and prospective applications are outlined and discussed, confirming paradigm's usefulness for solving hot world problems.

1 INTRODUCTION

To understand mental state of a handicapped person, problems of economy and ecology, or how to win on a battlefield, we must consider the system as a whole -- not just as a collection and interaction of parts. The situation may complicate dramatically if the system is dynamic and open, spreads over large territories, comprises unsafe or varying components, and cannot be observed in its entirety from a single point. Numerous world crises we have been witnessing at the beginning of this century, including the current economic one, may have emerged, first of all, due to our inability of seeing and managing complex systems as a whole.

To withstand the unwanted events and their consequences (ideally: predict and prevent them) we need effective worldwide integration of numerous efforts and often dissimilar and scattered resources and systems. Just establishing advanced communications between parts of the distributed systems and providing the possibility of sharing local and global information from any point, often called "interoperability", is becoming insufficient (even insecure and harmful) for solving urgent problems in dynamic environments, in real time and ahead of it.

We may need the whole distributed system to behave as an integral organism, with parts not so interoperating but rather complementing each other and representing altogether an integral whole pursuing global goals and having a sort of global awareness and consciousness. This whole should be essentially more than the sum of its parts, with the latter having sense, possibly even existence, in the context of this whole, rather than vice versa.

This paper develops further the over-operability principle researched in Sapaty, 1993, 1999, 2002, 2005 and other works (the term "over-operability" coined in Sapaty, 2002), which can establish intelligent dominant layer over distributed resources and systems, and help solve urgent world problems in a parallel, distributed, and dynamic way.

The rest of this paper compares the dominant atomistic approach in system design, implementation and management with holistic and gestalt principles, and describes a novel ideology and technology for integral solutions in distributed worlds, which can avoid many traditional management routines in solving global problems, with its numerous practical applications outlined and discussed.

2 ATOMISM, HOLISM, GESTALT

We used to exercise predominantly atomistic, parts-to-whole philosophy of the system design, comprehension and implementation, which extends even to the organization of management facilities themselves -- as a collection of interacting parts, or *agents*. (This philosophy actually being the same as a century ago.)

Originally a system or campaign idea and the functionality needed emerge in a very general form (in a single human mind or in a close collective of such minds). Then this general idea (shown symbolically in Fig. 1*a*) is partitioned into individual chunks, or "atoms", each detailed and studied further (Fig. 1*b*). This logical partitioning already causes swelling of the problem complexity (as indicated in Fig. 1*b*).

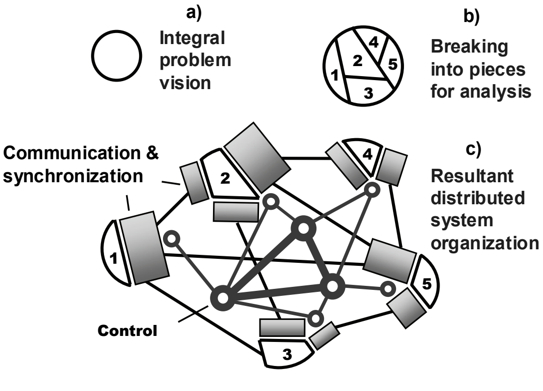

a)
Integral problem vision
b)
Breaking into pieces for analysis
1
2
3
4
5
Communication & synchronization
c)
Resultant distributed system organization
Control

Figure 1: System overhead under atomistic organization.

The next step is materialization of the defined parts and their distribution in physical or virtual space. To make these parts work or behave together within the original idea of Fig. 1*a*, we may need a good deal of their communication and synchronization, also sophisticated control infrastructures, as depicted in Fig. 1*c*. This overhead may be considerable, outweighing and shadowing the original project definition.

The main problem is that the initial idea (Fig. 1*a*) and even its second stage (Fig. 1*b*) are usually non formalized, remaining in the minds of creators only, and the *real system description and implementation start from the already partitioned-interlinked stage, with its huge overhead* (as Fig. 1*c*).

This parts-to-whole approach also dominates in the controversial "society of mind" theory (Minsky, 1988), which is trying to explain even human thinking from the atomistic positions.

Holism (see, for example, Smuts, 2007) has quite an opposite vision of systems:

- Holism as an idea or philosophical concept is diametrically opposed to atomism.
- Where the atomist believes that any whole can be broken down or analyzed into its separate parts and the relationships between them, the holist maintains that the whole is primary and often greater than the sum of its parts.
- The atomist divides things up in order to know them better; the holist looks at things or systems in aggregate.

Gestalt theory (Koffka, 1913; Wertheimer, 1922) is based on the holistic principles too:

- For the gestaltists, "Gestalten" are not the sums of aggregated contents erected subjectively upon primarily given pieces.
- Instead, we are dealing with wholes and whole–processes possessed of inner intrinsic laws.
- *Elements* are determined as parts by the intrinsic conditions of their wholes and are to be understood *as parts* relative to such wholes."

Although gestalt psychology and theory was a general approach, most of the work on gestalt was done in the area of perception. *In our research, we are trying to use the holistic and gestalt principles for the organization of distributed systems with highest possible integrity and performance* (see Sapaty, 2009).

3 WAVES, FIELDS, SCENARIOS

We describe here a novel organizational philosophy and model, based on the idea of spreading *interdependent parallel waves* (as shown in Fig. 2), as an alternative to the dominant atomistic approach briefed above, also under the influence of mentioned holistic and gestalt ideas.

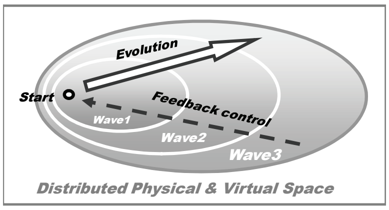


Figure 2: Grasping the entirety with spatial waves.

It allows us for an integral, parallel, and seamless navigation and coverage of virtual, physical or combined spaces where the solutions need to be found. Atomism emerges on the automatic implementation level only, which allows us to get high-level formal semantic definitions of systems and global operations in them, while omitting numerous organizational details (shown in Fig. 1*c*) and concentrating on global goals and overall performance instead.

An automatic materialization of this approach is carried out by the network of universal intelligent modules (U), embedded into important system points, which collectively interpret integral mission scenarios expressed in the waves formalism, which can start from any U, subsequently covering the distributed system at runtime.

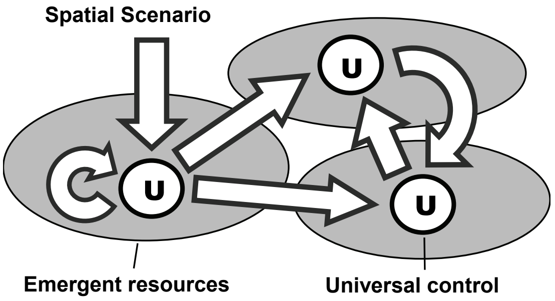


Figure 3: Self-spreading mission scenarios.

The wavelike scenarios are usually very compact and can be created and modified on the fly. They can cooperate or compete with each other in the distributed networked space as overlapping fields of parallel solutions.

Spreading waves can create knowledge infrastructures arbitrarily distributed between system components (robots, sensors, humans). These, subsequently or simultaneously navigated by same or other waves, can effectively support distributed databases, command and control, situation awareness, and autonomous decisions.

This paradigm is much in line with the existing abundant evidence that certain aspects of cognition, morals, needs, object relations, motor skills, and language acquisition proceed in developmental stages. These stages appear to be fluid, flowing, overlapping waves (Wilber, 2009), where also:

- Each stage has a holistic pattern that blends all of its elements into a structured whole;
- These patterns unfold in a relational sequence, with each senior wave transcending but including its juniors.

Our approach is also consistent with the ideas of self-actualization and person-centered approach (Rogers, 1978; Kriz, 2008), where the self is considered as an organized, consistent, conceptual gestalt exhibiting active forward thrust -- against tension reduction, equilibrium, or homeostasis (as in Freud, 2007, and others). In our case, instead of a single person we have the whole distributed system with high integrity and “active global thrust” behavior.

4 THE SCENARIO LANGUAGE

Distributed Scenario Language, or DSL (and its previous versions, WAVE including, as in Sapaty, 1999, 2005) reflects the waves model proposed, and allows us to directly express semantics of problems to be solved in distributed worlds, also the needed global system behavior in a non-atomistic manner. DSL operates with:

- *Virtual World (VW)*, which is discrete and consists of nodes and links connecting these nodes.
- Continuous *Physical World (PW)*, any point in which may be accessed by physical coordinates (taking into account certain precision).
- *Virtual-Physical World (VPW)*, which is an extension of VW where nodes additionally associate with certain coordinates in PW.

It also has the following key features:

- A DSL scenario develops as a transition between sets of progress points (or *props*) in the form of parallel *waves*.
- Starting from a prop, an action may result in one or more props (the resultant set of props may include the starting prop too).
- Each prop has a resulting *value* (which can be multiple) and a resulting *state* (being one of the four: *thru*, *done*, *fail*, and *abort*).
- Different actions may evolve independently or interdependently from the *same* prop, contributing to (and forming altogether) the resultant set of props.
- Actions may also *spatially succeed each other*, with new ones applied in parallel from all the props reached by preceding actions.
- Elementary operations can directly use local or remote values of props obtained from other actions (or even from the whole scenarios).
- Elementary operations can result either in open values that can be directly used as *operands* by other operations in an expression, or by the *next*

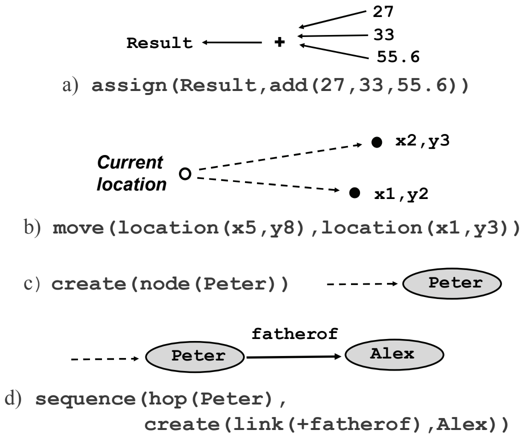
operations in a sequence. They can also be directly assigned to *local or remote variables* (for the latter case, an access to these variables may invoke scenarios of any complexity).

- Any prop can associate with a *node* in VW or a *position* in PW, or *both* -- when dealing with VPW.
- Any number of props can be simultaneously linked with the same points of the worlds.
- Staying with world points (virtual, physical, or combined) it is possible to *directly access* and update local data in them.
- Moving in physical, virtual or combined worlds, with their possible modification or even creation from scratch, are as routine operations as, say, arithmetic or logical operations of traditional programming languages.
- DSL can also be used as a usual universal programming language (like C, Java, or FORTRAN).

DSL has a recursive syntax, which on top level is as follows:

wave	→	***phenomenon* | *rule* ({ *wave* , })**
phenomenon	→	***constant* | *variable* | *special***
constant	→	***information* | *matter* | *combined***
variable	→	***heritable* | *frontal* | *environmental* | *nodal***
rule	→	***movement* | *creation* | *elimination* | *echoing* | *fusion* | *verification* | *assignment* | *advancing* | *branching* | *transference* | *timing* | *granting***

Elementary programming examples in DSL are shown in Fig. 4 for: a) assignment of a sum of values to a variable; b) parallel movement into two physical locations; c) creation of a node in a virtual space, and d) extension of the latter with a new link and node.



a) `assign(Result,add(27,33,55.6))`

b) `move(location(x5,y8),location(x1,y3))`

c) `create(node(Peter))`

d)
```
sequence(hop(Peter),
         create(link(+fatherof),Alex))
```

Figure 4: Elementary examples in DSL.

Traditional abbreviations of operations and delimiters can also be used, as in many further examples throughout this text, to simplify and shorten DSL programs, remaining however within the general recursive syntactic structure shown above.

5 COMPOSITION OF WAVES

The language allows for an integral parallel navigation of distributed worlds in a controlled *depth and breadth mode*, with any combinations of the two. We will highlight here key possibilities of doing this by composition of DSL scenarios, or waves.

5.1 Single Wave Features

Single wave (let it be *W1*) development features are shown in Fig. 5. Starting from a prop, which may be associated with a point in the world, the related scenario evolves, grasps, and covers certain region in it, performing any operations needed in the distributed space.

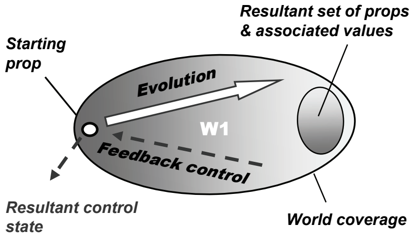
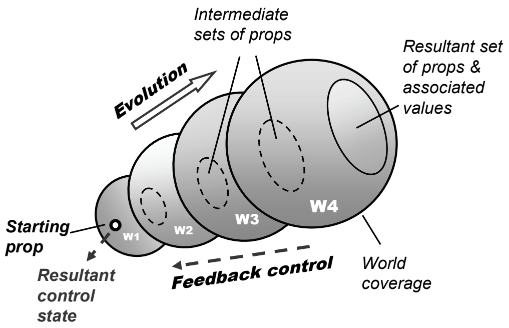


Figure 5: Single wave features.

The result of this spatial evolution may be multiple, and may lie in a (final) sub-region of the region covered, being represented by a set of resultant props (each linked to world points) and associated with them values. After termination of the wave, its resultant control state (which, in a parallel feedback process, merges termination states throughout the region covered) is available in the starting prop, and may be taken into account for decisions at higher levels. Also, if requested from higher levels, the values associated with the resultant props (which may be remote) can be lifted, spatially raked, and returned to the starting prop for a further processing.

5.2 Advancing in Space

The depth mode development of waves is shown in Fig. 6. For this type of composition, each subsequent wave is applied in parallel from all props in space reached by the previous wave, with the resultant set of props (and associated values) on the whole group being the one of the last applied wave (i.e. *W4* in the figure).



`advance(W1,W2,W3,W4)`

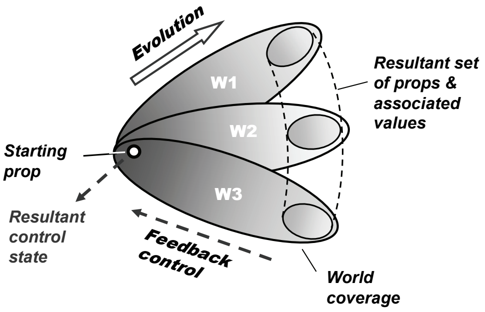
Figure 6: Depth mode composition of waves.

This spatial advancement of waves returns the resultant control state which is available at the starting prop, and the values of the resultant set of props can also be echoed to the starting prop if requested. Examples of other advancing rules:

- `advance synchronized` – the one where any new wave is applied only after all invocations of the previous wave have been terminated;
- `repeat` – where the same wave is applied repeatedly from all props reached by its previous invocation;
- `repeat synchronized` – where in the repeated invocation of a wave each new invocation starts only after full completion of the previous one.

5.3 Branching in Space

The branching breadth mode composition of waves is shown in Fig. 7, where all waves in the group are evolving from the same starting prop, and each wave, with its own resultant set of props and associated values, contributes to the final result.



`sequence(W1,W2,W3)`

Figure 7: Breadth mode composition of waves.

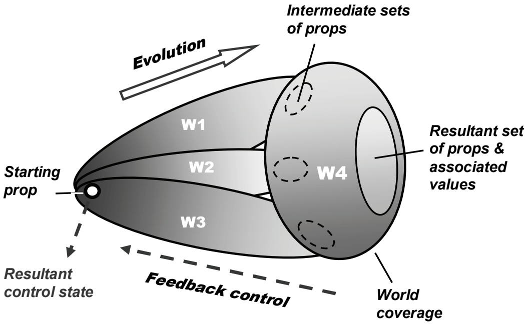
The merge of results from different waves depends on the branching rule used, with their repertoire (besides the `sequence` in Fig. 7) including:

`if, while, parallel, or, parallel or, and, parallel and, cycle, loop, and sling.`

(More details on these and other rules can be found, say, from Sapaty, 1999, 2005.)

5.4 Combined Branching-Advancing

Any combination of advancing and branching modes in a distributed space can be expressed and implemented in DSL (as shown in Fig. 8).



`advance(sequence(W1,W2,W3),W4)`

Figure 8: Breadth–depth composition mode.

These combinations, when embraced by the existing variety of composition rules, can provide any imaginable and even so far unimaginable spatial algorithms that can solve distributed problems in highly integral and compact ways, without explicit descending to the traditional atomistic level shifted to the automatic implementation only.

5.5 Operations on Remote Values

Due to fully recursive organization of DSL, it is possible to program in it arbitrary complex expressions directly operating not only on local but also arbitrarily remote values, where any programs (scenarios) can happen to be operands of any operations (expressed by rules). This gives an enormous expressive power and compactness to complex spatial scenarios evolving in distributed environments. An example of such compact expression of spatial operations on remote values and variables is shown in Fig. 9.

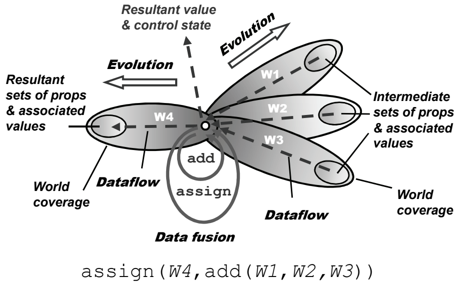

Resultant value & control state
Evolution
Evolution
Resultant sets of props & associated values
W1
W2
W3
W4
Intermediate sets of props & associated values
add
assign
World coverage
Dataflow
Dataflow
World coverage
Data fusion
assign(W4,add(W1,W2,W3))

Figure 9: Direct operations on remote values.

6 DISTRIBUTED INTERPRETER

DSL interpreter, as from the previous language version called WAVE (Sapaty, 1993, 1999, 2005), has been prototyped in different countries on various platforms. Its public domain version (financed in the past by Siemens/Nixdorf) is being used for applications like intelligent network management or simulation of distributed dynamic systems. The DSL interpreter basics include:

- It consists of a *number of specialized modules* working in parallel and handling and sharing specific data structures, which are supporting persistent virtual worlds and temporary hierarchical control mechanisms.
- The whole network of the interpreters can be *mobile and open*, changing at runtime the number of nodes and communication structure between them.
- The heart of the distributed interpreter is its *spatial track system* enabling hierarchical command and control and remote data and code access, with high integrity of emerging parallel and distributed solutions.

The DSL interpreter structure is shown in Fig. 10.

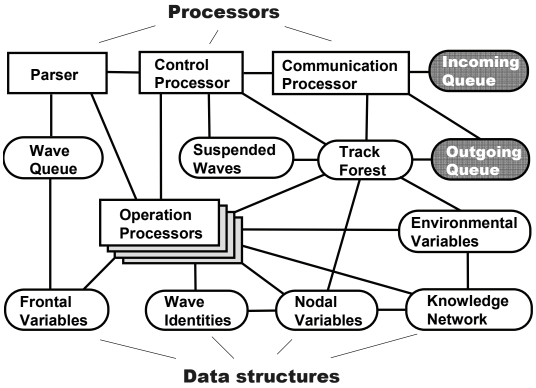


Figure 10: Structure of DSL interpreter.

It can be easily implemented in both software and hardware on any platforms, where the intelligent "wave chip" can be implanted into a great variety of devices, making them working together as an integral unit under the spatial DSL scenarios.

7 PROGRAMMING EXAMPLES

We will show here examples of solution in DSL of some important problems on networks and graphs in a fully distributed way, where each node may reside in a separate computer.

7.1 Shortest Paths

The solution for finding a path between two nodes by navigating the network with parallel waves is sown in Fig. 11, and the scenario that follows.

```
sequence(
 (direct # a; Ndist = 0; repeat(
  any #; Fdist += LINK;
  Ndist == nil, Ndist > Fdist;
  Ndist = Fdist; Npred = BACK))
 (direct # e; repeat(
  Fpath &= CONT; any # Npred);
  USER = Fpath))
```

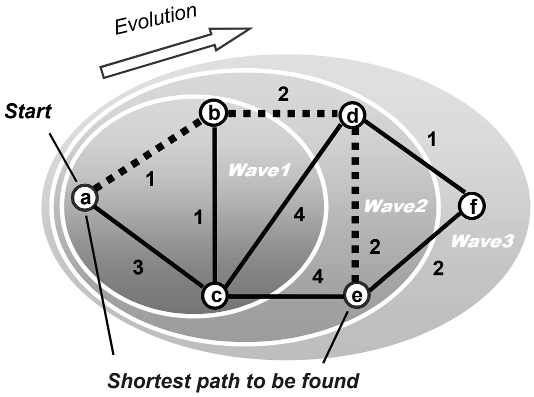


Figure 11: Finding shortest path with waves.

Many problems of optimization and control may be expressed as finding shortest paths in a distributed solution space.

7.2 Spatial Topology Analysis

DSL allows us to directly analyze and process distributed topologies in a parallel and extremely concise way.

7.2.1 Articulation Points

To find the weakest nodes in a network (called *articulation points*) which, when removed, split it into disjoint parts, as in Fig. 12 for node d, we need only the program that follows.

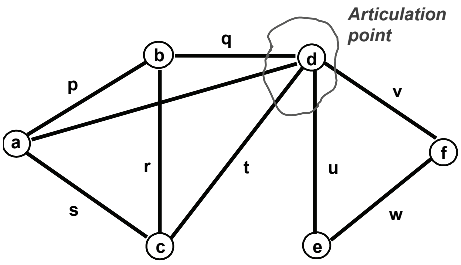


Figure 12: Articulation points.

```
direct # all; ID = CONT; Nm = 1;
and((random(all #);
 repeat(Nm ==; Nm = 1; all #)),
 (all #; Nm ==), USER = CONT)
```

Result: d.

7.2.2 Cliques

Cliques (or fully connected sub-graphs of a graph, as in Fig. 13), on the opposite, may be considered as strongest parts of a system. They can be found in parallel by the program that follows.

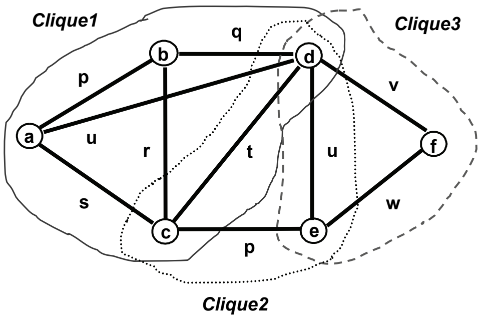


Figure 13: Cliques.

```
direct # all; Fclique = CONT;
repeat(all #; CONT !~ Fclique;
 and(andpar(any # Fclique; done !),
  or((BACK > CONT; done !),
   Fclique &= CONT))); USER = Fclique
```

Result: (a,b,c,d), (c,d,e), (d,e,f)

7.2.3 All Triangles

Any topological patterns can be found in any distributed network. For example, finding all triangles in a graph in Fig. 13 needs a simple code:

```
direct # all; Ftr = CONT;
2(all#; BACK > CONT; Ftr &= CONT);
 any # Ftr : 1; USER = Ftr
```

Result: (a,b,c), (b,c,d), (c,d,e), (d,e,f), (a,b,d), (a,c,d)

7.2.4 Network Creation

Any network can be created in a distributed space, and in parallel mode, by a very simple code too, as follows, as for the network in Fig. 13.

```
create(direct#a; p#b; q#d; u##a,(v#f;
w#e; u##d,(p#c; s##a, r##b, t##d)))
```

Arbitrary infrastructures can be created at runtime, on the fly, which can become active by putting certain procedures into their nodes and links. Any other existing models (incl. Petri nets, neural nets, contract nets, etc.) can also be implemented in a fully distributed and parallel way in DSL. Many related examples can be found in Sapaty, 1999.

8 COLLECTIVE ROBOTICS

Installing DSL interpreter into mobile robots (ground, aerial, or underwater) may allow us to organize any group solutions of complex problems in distributed physical spaces in a concise and effective way, shifting traditional management routines to automatic level. It is possible to express tasks and behaviors on different levels, as follows.

8.1 Task Level

Heterogeneous groups of mobile robots (as in Fig. 14) can be tasked at a highest possible level, just telling what they should do together, without detailing how, and what are the duties of every unit. An example task may be formulated as follows.

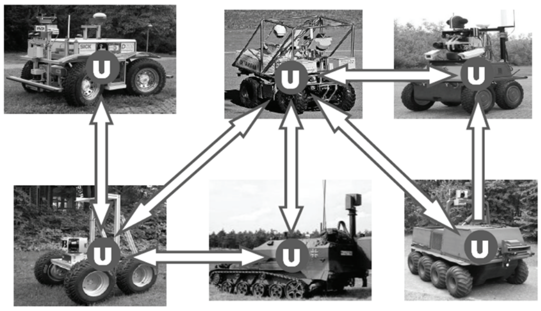


Figure 14: Grouping ground vehicles.

Go to physical locations of the disaster zone with coordinates (50.433, 30.633), (50.417, 30.490), and (50.467, 30.517). Evaluate damage in each location, find and transmit the maximum destruction value, together with exact coordinates of the corresponding location, to a management center.

The DSL program will be as follows:

```
transmit(maximum(
  move((50.433, 30.633),
       (50.417, 30.490),
       (50.467, 30.517));
  evaluate(destruction)& WHERE))
```

Details of automatic implementation of this scenario by different numbers of mobile robots are discussed in (Sapaty, 2009c).

8.2 Behavioral Level

After embedding DSL interpreters into robotic vehicles (like the aerial ones in Fig. 15), we can also provide any needed detailed collective behavior of them (at a lower than top task level, as before)—from *loose swarms* to a *strictly controlled integral unit* obeying external orders. Any mixture of different behaviors within the same scenario can be easily programmed too.

The following DSL scenario combines loose, random swarm movement in a distributed space with periodic finding/updating topologically central unit, and setting runtime hierarchical infrastructure between the units. The latter controls observation of distributed territory, collecting potential targets, distributing them between the vehicles, and then impacting potential targets by them individually. More on the implementation of this scenario can be found in Sapaty, 2008.

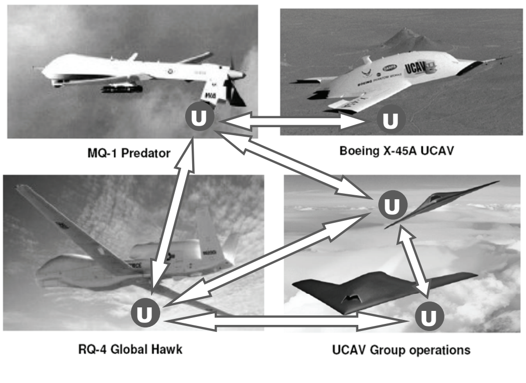


Figure 15: Grouping aerial vehicles.

```
(hop(allnodes); Range = 500;
 Limits = (dx(0,8), dy(-2,5));
 repeat(Shift = random(Limits);
  if(empty(hop(Shift, Range),
     move(Shift)))),
(repeat(hop(
   Faver =average(hop(allnodes);WHERE);
   min(hop(allnodes);
   distance(Aver, WHERE)& ADDRESS):2));
   stay(hop(nodes,all);rem(links,all);
   Frange = 20; repeat(
    linkup(+infra, firstcome, Frange));
   orpar(
    loop(nonempty(Fseen =
     repeat(free(detect(targets)),
     hoplinks(+ infra));
     repeat(
      free(select_move_shoot(Fseen),
      hoplinks(+ infra))),
    sleep(360)))
```

9 OTHER APPLICATIONS

9.1 Distributed Avionics

Distributed communicating DSL Interpreters, embedded into aircraft's key mechanisms (as in Fig. 16), can provide highest possible integrity of the aircraft that may continue to *function as a whole* even under *physical disintegration* -- which may help find critical runtime solutions saving lives and equipment (see also Sapaty, 2008a).

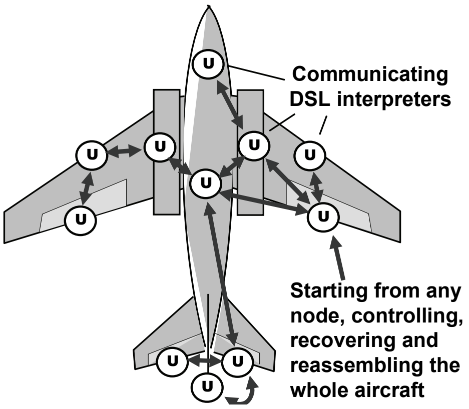

Communicating DSL interpreters
U
Starting from any node, controlling, recovering and reassembling the whole aircraft

Figure 16: Distributed control infrastructure.

Collecting availability of aircraft's basic mechanisms, and establishing overall aircraft control from any available DSL interpreter, may be organized as follows:

```
Available =
 repeat(free(belong(CONT,
  (left_aileron, right_aileron,
   left_elevator,right_elevator,
   rudder, left_engine,right_engine,
   left_chassis, right_chassis, …));
   CONT), hop(firstcome, neighbors));
if(sufficient(Available),
   control(Available), set(alarm))
```

9.2 Objects Tracking

In a large distributed space, each embedded (or moving) sensor can handle only a limited part of space, so to keep the whole observation continuous, the mobile object seen should be handed over between neighboring sensors during its movement, along with the data accumulated on it (see also Sapaty, 1999, 2007, 2008).

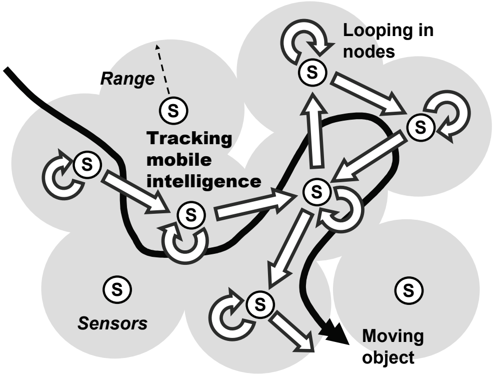


Figure 17: Tracking mobile objects.

The following program, starting in all sensors, *catches the object* it sees and *follows* it wherever it goes, if not observable from this point any more.

```
hop(allnodes); Fthr = 0.1;
Fobj = search(aerial);
visibility(Fobj) > Fthr; repeat(
 loop(visibility(Fobj) > Fthr);
 maxdest(hop(neighbors); (Seen =
 visibility(Fobj)) > Fthr; Seen))
```

9.3 Emergency Management

Embedded communicating DSL Interpreters can convert any post-disaster wreckage into a universal spatial machine capable of self-analysis and self-recovery under integral management scenarios (as in Sapaty, Sugisaka, Finkelstein, Delgado-Frias, Mirenkov, 2006; Sapaty, 2006). For example, all casualties counting program may be as follows (with its distributed operation shown in Fig. 18):

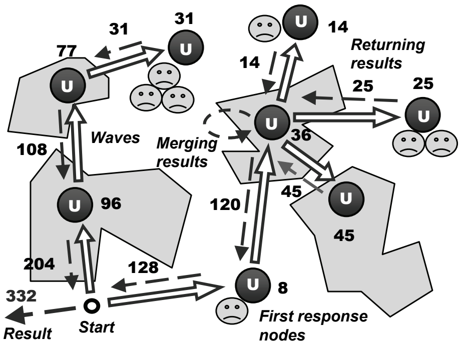


Figure 18: Counting all casualties.

```
Farea = disaster area definition;
output(sum(hop(Farea);
 repeat(free(count(casualties)),
 hop(alllinks, firstcome, Farea))))
```

Counting casualties in each region separately and organizing proportional relief delivery to each of them, may be expressed as follows:

```
Frea = disaster area definition;
split(collect(hop(Farea));
 repeat(done(count(casualties)&WHERE),
   hop(anylinks, firstcome, Farea))));
Fsupply = replicate(“package”, VAL:1);
move(VAL:2); distribute(Fsupply)
```

9.4 Directed Energy Systems

Directed energy systems and weapons are of rapidly growing importance in many areas, and especially in critical infrastructure protection, also on advanced battlefields (as shown in fig. 19). With the hardware equipment operating with the speed of light, traditional manned C2 is becoming a bottleneck for these advanced technical capabilities. With the technology offered, we may organize any runtime C2 infrastructures operating automatically, with the "speed of light" too, fitting the hardware capabilities and excluding men from the loop in time critical situations.

Figure 19: DEW in an advanced battlespace.

The following is an example of setting an automatic runtime C2 in a system with direct energy (DE) source, relay mirror (RM), and a target discovered, with an operational snapshot shown in Fig. 20.

```
sequence(
  parallel(
    (hop(DE); adjust(RM)),
    (hop(RM); adjust(DE, Target))),
  (hop(DE); activate(DE)))
```

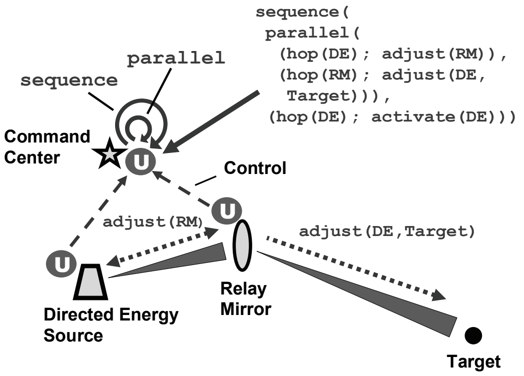


Figure 20: DE-RM-target operational snapshot.

There also exist advanced projects of global dominance with transference of directed energy, like the Boeing's Advanced Relay Mirror System (ARMS) concept. It plans to entail a constellation of as many as two dozen orbiting mirrors that would allow 24/7 coverage of every corner of the globe. When activated, this would enable a directed energy response to critical trouble spots anywhere.

We can use the distributed shortest path solution shown in section 7.1 for providing a runtime path in a worldwide distributed *dynamic* set of relay mirrors (as some of which may happen to be out of order) -- between the DE source and destination needed. This will enable optimal directed energy transfer, as shown in Fig. 21 (see also Sapaty, Morozov, Sugisaka, 2007).

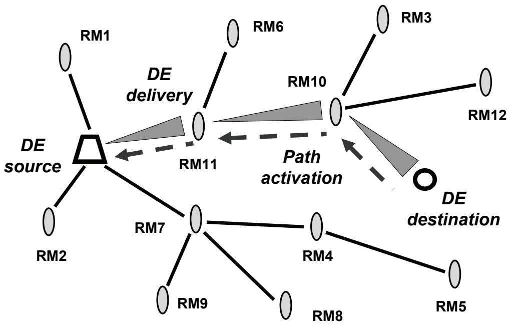


Figure 21: DE delivery via network of relay mirrors.

9.5 Electronic Warfare

Often the picture in Fig. 22 is shown as a typical example of electronic warfare. But this may rather be the last chance to survive from a missile attack. Involvement of many diverse and interlinked systems, especially for preventing and anticipating the attacks, which may be multiple and simultaneous, should be of paramount importance. All existing and being developed electronic support, attack, and protection measures have very limited scope and effect if used alone. But taken together they may provide a capability for fulfilling the objectives required. And the technology offered can readily organize this (as in Sapaty, 2007a, 2009a).

Figure 22: A Lockheed plane releasing decoy flares.

Instead of *physical flares* thrown from a plane in the final moments, we may throw, throughout the region in danger, which may be worldwide, the "*DSL scenario flares"* that can dynamically unite any available DE facilities and systems in an overwhelming electronic response to any threats.

9.6 Robotized Armies

Distributed robotized systems are of rapidly growing importance in defense (Singer, 2009, 2009a), where robotic swarming on asymmetric battlefields is becoming a major dimension of the new military doctrine for 21st century. But, as admitted by Singer, 2009, swarming, along with its simple rules of individual behavior and fully distributed nature, agility, and ubiquity, may also result in unpredictability of behavior for both sides of the conflict.

The approach briefed in this paper, also investigated in previous publications on this paradigm, is very much in line with these modern trends. Moreover, we are offering a unified solution that can harness loosely coupled swarms, always guaranteeing their global-goal-driven behavior, where the watershed between loose swarming and strict hierarchical control may be situation dependent and changing over time (as programmed in Section 8.2).

These new doctrine trends will inevitably influence the role and sense of communications on battlefields, as with the planned drastic reduction of centralized C2 much more emphasis will be paid to intelligent tactical communications, where the scenario mobility in networked systems, offered by the approach proposed, may constitute an effective solution, with the key points (as in Sapaty, 2009b):

- Dramatic shift of global organization to intelligent tactical communications;
- Self-spreading and self-recovering mission scenarios and emergent command and control;
- Embedding intelligent protocol module into existing communication equipment;
- Situation-dependent watershed between global control and local communications.

In relation to the said above, different (including new) types of commands and control strategies for distributed robotized systems were investigated in DSL (Sapaty, Morozov, Sugisaka, Finkelstein, Lambert, 2008).

10 THE FIRST COMPUTERS

The approach offered may be compared with the invention of the first world computers (Rojas, 1997) and first high-level programming languages (Zuse, 1948/49). In our case, this computer may not only operate with data stored in a localized memory, but can cover, grasp, and manage any distributed system, the whole world including, and can work not only with information but with physical matter or objects too.

If compared with the Turing computational model, instead of the head moving through tape in performing calculations, we have a recursive formula that unwraps, replicates, covers and matches the distributed world in parallel, scanning it forth and back, bringing operations and data directly to the points of their consumption, automatically setting distributed command and control infrastructures, and organizing local and global behaviors needed.

The term "computer" first referred to the people who did scientific calculations by hand (Grier, 2005). In the end, they were rewarded by a new electronic machine that took the place and the name of those who were, once, the computers.

We can draw the following symbolic parallel with this. Despite the overwhelming automation of human activity (in both information and matter processing) the world as a whole may still be considered as remaining a *human machine*, as main decisions and global management still remain the human prerogative.

With the approach offered, we can effectively automate this top-level human supervision, actually converting the whole world into a *universal programmable machine* spatially executing global scenarios in DSL or a similar language. Despite certain science fiction flavor of this comparison, we can find numerous applications for such a global approach, some mentioned above, where top level decision automation could withstand unwanted events and save lives, and where timely human reaction may not be possible, even in principle.

11 CONCLUSIONS

We have developed and tested a novel system approach, which can describe what the system should do and how to behave on a higher level, while delegating traditional management details (like partitioning into components, their distribution, interaction and synchronization) to the effective automatic layer.

A DSL scenario is not a usual program -- it is rather a recursive active spatial pattern dynamically matching structures of distributed worlds. It has a hierarchical organization, which is grasping, by means of spreading parallel waves, the whole of the system to be comprehended and impacted.

The DSL scenarios can also create, in a parallel and fully distributed way, active distributed worlds, which become persistent and operate independently. They may spatially intervene into operation of these and other worlds and systems, changing their structures and behaviors in the way required, also self-recover from indiscriminate failures and damages, as well as repair and recover the systems managed.

Prospective applications of this work can also be linked with economy, ecology and weather prediction—by using the whole networked world as a spatial supercomputer, self-optimizing its performance. Also, for terrorism and piracy fight, where the powerful parallel ability of analyzing distributed systems and finding strong and weak patterns in them, as well as any structures (as shown in Section 7.2) may be the key to global solutions.

Crises may spark anywhere and anytime like, say, birds or swine flu or the current global economic disaster. We must be ready to react on them quickly and asymmetrically, withstanding and eradicating them -- in a "pandemic" way too, highly organized and intelligent, however.

We already have technical capabilities for this, as for example, the number of mobile phone owners in the world is approaching 3bn, and installing DSL interpreter in at least a fraction of them, can allow us to organize collective runtime (and ahead of it) response to any world events.

ACKNOWLEDGEMENTS

This work has been sponsored by the Alexander von Humboldt Foundation in Germany. Special thanks to Anatoly Morozov, Vitaly Klimenko, Yuri Korovitsky, Vladimir Voloboev, Eugene Bondarenko, and Anatoly Belyaev from the National Academy of Sciences of Ukraine for years of invaluable support and productive, often hot, discussions of these ideas. Recent chats with Juergen Kriz at the 16th GTA Convention in Osnabrueck, Germany, strengthened author's admiration of the gestalt psychology and theory, influencing general orientation of this paper.

REFERENCES

Freud, S., 1997. *General Psychological Theory*. Touchstone.

Grier, D. A., 2005. When Computers Were Human, Princeton University Press.

Koffka, K., 1913. Beiträge zur Psychologie der Gestalt. *von F. Kenkel. Zeits. f.Psychol*., 67.

Kriz, J., 2008. *Self-Actualization: Person-Centred Approach and Systems Theory*. PCCS Books.

Minsky, M., 1988. *The Society of Mind.* Simon and Schuster, New York.

Rojas, R., 1997. Konrad Zuse's Legacy: The Architecture of the Z1 and Z3. *IEEE Annals of the History of Computing*, Vol. 19, No. 2.

Rogers, C. R., 1978. *Carl Rogers on Personal Power: Inner Strength and Its Revolutionary Impact*. Trans-Atlantic Publications.

Sapaty P., 2009. Gestalt-Based Ideology and Technology for Spatial Control of Distributed Dynamic Systems, *Proc. International Gestalt Theory Congress, 16th Scientific Convention of the GTA*. University of Osnabrück, Germany.

Sapaty, P., 2009a. Distributed Capability for Battlespace Dominance. In *Electronic Warfare 2009 Conference & Exhibition*, Novotel London West Hotel & Conference Center, London.

Sapaty, P., 2009b. High-Level Communication Protocol for Dynamically Networked Battlefields. *Proc. International Conference Tactical Communications 2009 (Situational Awareness & Operational Effectiveness in the Last Tactical Mile)*. One Whitehall Place, Whitehall Suite & Reception, London, UK.

Sapaty, P. S., 2009c. Providing Spatial Integrity For Distributed Unmanned Systems". *Proc. 6th International Conference in Control, Automation and Robotics ICINCO 2009*. Milan, Italy.

Sapaty, P., 2008. Distributed Technology for Global Dominance. *Proc. of SPIE* -- Volume 6981, Defense Transformation and Net-Centric Systems 2008, Raja Suresh, Editor, 69810T.

Sapaty, P., 2008a. Grasping the Whole by Spatial Intelligence: A Higher Level for Distributed Avionics. *Proc. International Conference Military Avionics 2008*, Cafe Royal, London, UK.

Sapaty, P., Morozov, A., Finkelstein, R., Sugisaka, M., Lambert, D., 2008. A new concept of flexible organization for distributed robotized systems. *Artificial Life and Robotics*, Volume 12, Nos. 1-2/ March, Springer Japan.

Sapaty, P., 2007. Intelligent management of distributed sensor networks. In *Sensors, and Command, Control, Communications, and Intelligence (C3I) Technologies for Homeland Security and Homeland Defense VI*, edited by Edward M. Carapezza, *Proc. of SPIE*, Vol. 6538, 653812.

Sapaty, P., 2007a. Global Management of Distributed EW-Related Systems. *Proc. International Conference Electronic Warfare: Operations & Systems*. Thistle Selfridge, London, UK.

Sapaty, P., Morozov, A., Sugisaka, M., 2007. DEW in a Network Enabled Environment. *Proc. International Conference Directed Energy Weapons 2007*. Le Meridien Piccadilly, London, UK.

Sapaty, P., 2006. Crisis Management with Distributed Processing Technology. *International Transactions on Systems Science and Applications*. Vol. 1, no. 1.

Sapaty, P., Sugisaka, M., Finkelstein, R., Delgado-Frias, J., Mirenkov, N., 2006. Advanced IT Support of Crisis Relief Missions. *Journal of Emergency Management.* Vol.4, No.4, July/August.

Sapaty, P. S., 2005. *Ruling Distributed Dynamic Worlds*. John Wiley & Sons, New York.

Sapaty, P. S., 2002. Over-Operability in Distributed Simulation and Control. *The MSIAC's M&S Journal Online*. Winter Issue, Volume 4, No. 2, Alexandria, VA, USA.

Sapaty, P. S., 1999. *Mobile Processing in Distributed and*

Sapaty, P. S., 1993. A distributed processing system, *European Patent No. 0389655*. European Patent Office.

Singer, P. W., 2009. Wired for War. Robots and Military Doctrine. *JFQ / issue 52*, 1st quarter.

Singer, P. W., 2009a. *Wired for War: The Robotics Revolution and Conflict in the 21st Century*. Penguin.

Smuts, J. C., 2007. *Holism And Evolution*. Kessinger Publishing, LLC

Wertheimer, M., 1924. *Gestalt Theory*. Erlangen, Berlin.

Wilber, K. 2009. *Ken Wilber Online: Waves, Streams, States, and Self--A Summary of My Psychological Model (Or, Outline of An Integral Psychology)*. Shambhala Publications.

Zuse, K., 1948/49. Uber den Plankalk, als Mittel zur Formulierung schematisch kombinativer Aufgaben. In *Archiv Mathematik*, Band I.

BRIEF BIOGRAPHY

Dr Peter Simon Sapaty, chief research scientist and director of distributed simulation and control project at the Institute of Mathematical Machines and Systems, National Academy of Sciences of Ukraine, is with networked systems for more than four decades. A power network engineer on education, he created citywide heterogeneous computer networks from the end of the sixties, implemented a multiprocessor macro-pipeline supercomputer in the seventies-eighties, and since then used distributed computer networks for solving complex problems of most different natures—from distributed knowledge bases to intelligent network management to road traffic control to simulation of battlefields. He also worked in Germany, UK, Canada, and Japan as Alexander von Humboldt Foundation fellow, project leader, research professor, department head, and special invited professor; created and chaired a SIG on mobile cooperative technologies within Distributed Interactive Simulation project in the US. Peter invented a higher-level distributed networking technology used in different countries and resulted in a European Patent and two John Wiley books. His interests include coordination and simulation of large distributed dynamic systems under the holistic and gestalt principles, with application in advanced command and control, cooperative robotics, infrastructure protection, crisis management, and especially for finding asymmetric solutions in unpredictable and hostile environments.

SIGNAL PROCESSING, SYSTEMS MODELING AND CONTROL

Full Papers

INFORMATION-THEORETIC VIEW OF CONTROL

Prateep Roy, Arben Çela
Université Paris-Est, ESIEE, Département Systèmes Embarqués, Paris, France
royp@esiee.fr, celaa@esiee.fr

Yskandar Hamam
F'SATIE-TUT, Pretoria, South Africa
hamama@tut.ac.za

Keywords: Information Theory, Shannon Entropy, Mutual-Information, Control, Bode Sensitivity Integral.

Abstract: In this paper we are presenting the information-theoretic explanation of Bodé Sensitivity Integral, a fundamental limitation of control theory, controllability grammian and the issues of control under communication constraints. As resource-economic use of information is of prime concern in communication-constrained control problems, we need to emphasize more on informational aspect which has got direct relation with uncertainties in terms of Shannon Entropy and Mutual Information. Bode Integral which is directly related to the disturbances can be correlated with the difference of entropies between the disturbance-input and measurable output of the system. These disturbances due to communication channel-induced noises and limited bandwidth are causing the information packet-loss and delays resulting in degradation of control performances.

1 INTRODUCTION

In recent years, there has been an increased interest for the fundamental limitations in feedback control. Bode's sensitivity integral (Bode Integral, in short) is a well-known formula that quantifies some of the limitations in feedback control for linear time-invariant systems. In (Sandberg and Bernhardsson, 2005), it is shown that there is a similar formula for linear time-periodic systems.

In this paper, we focus on Bode integral of control theory and Shannon Entropy of information theory because the latter is a stronger metric for uncertainty which hinders control of a system.

It has been known that control theory and information theory share a common background as both theories study signals and dynamical systems in general. One way to describe their difference is that the focal point of information theory is the signals involved in systems while control theory focuses more on systems which represent the relation between the input and output signals. Thus, in a certain sense, we may expect that they have a complementary relation. For this reason, many researchers have consecrated studies on the interactions of the two theories : Control Theory and Information Theory.

In networked control systems, there are issues related to both control and communication since communication channels with data losses, time delays, and quantization errors are employed between the plants and controllers (Antsaklis and Baillieul, 2007). To guarantee the overall control performance in such systems, it is important to evaluate the quantity of information that the channels can transfer. Thus, for the analysis of networked control systems, information theoretic approaches are especially useful, and notions and results from this theory can be applied. The results in (Nair and Evans, 2004) and (Tatikonda and Mitter, 2004) show the limitation in the communication rate for the existence of controllers, encoders, and decoders to stabilize discrete-time linear feedback systems.

The focus of information theory is more on the signals and not on their input-output relation. Thus, based on information theoretic approaches, we may expect to extend prior results in control theory. One such result can be found in (Martins et al., 2007), where a sensitivity property is analyzed and Bode's integral formula (Bode H., 1945) is extended to a more general class of systems. A fundamental limitation of sensitivity functions is presented in relation to the unstable poles of the plants.

2 PROBLEM FORMULATION

Networked control systems suffer from the drawbacks of packet losses, delays and quantization in particular. These cause degradation of control performances and under some conditions instability. Uncertainties due to packet losses, delays, quantization, communication channel induced noises etc. have a great influence on the system systems performance. If we consider only the uncertainties induced by channel noise and quantization we may write:

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t) + w(t); \qquad (1)\\ y(t) &= Cx(t) + Du(t) + v(t); \end{aligned}$$

where $A \in R^{n\times n}$ is the system or plant matrix and $B \in R^{n\times q}$ is the control or input matrix. Also, $x(t)$ is the state, $u(t)$ is the control input, $y(t)$ is the output, C is the output or measurement matrix, D is the Direct Feed matrix, $w(t)$ and $v(t)$ are the external disturbances and noises of Gaussian nature respectively. Our aim is to achieve better control performance of system by tackling these uncertainties using Shannon's Mutual-Information, Information-Theoretic Entropy and Bode Sensitivity. We present the information-theoretic model of such uncertainties and their possible reduction using information measures.

3 PRELIMINARIES

By means of the connection between Bode integral and the entropy cost function, paper (Iglesias, 2001) provided a time-domain characterization of Bode integral. The traditional frequency domain interpretation is that, if the sensitivity of a closed-loop system is decreased over a particular frequency range typically the low frequencies the designer "pays" for this in increased sensitivity outside this frequency range. This interpretation is also valid for the time-domain characterization presented in (Iglesias, 2001) provided one deals with time horizons rather than frequency ranges. Time-domain characterization of Bode's integral shows how the frequency domain trade-offs translate into the time-domain. Under the usual connection between the time and frequency domains: low (high) frequency signals are associated with long (short) time horizons. In Bode's result, it is important to differentiate between the stable poles, which do not contribute to the Bode sensitivity integral and the unstable poles, which do. Time-varying systems which can be decomposed into stable and unstable components are said to possess an exponential dichotomy. What the exponential dichotomy says is that the norm of the projection onto the stable subspace of any orbit in the system decays exponentially as $t \rightarrow \infty$ and the norm of the projection onto the unstable subspace of any orbit decays exponentially as $t \rightarrow -\infty$, and furthermore that the stable and unstable subspaces are conjugate. The existence of an exponential dichotomy allows us to define a stability preserving state space transformation (a Lyapunov transformation) that separates the stable and unstable parts of the system.

3.1 Mutual Information

Shannon's Mutual information is just the information carried by one random variable about the other. It is a quantity in the time domain. Mutual Information $I(X;Y)$, between X as the input variable and Y as the output variable, has the lower and upper bounds given by the following:

$$R(D) = RateDistortion = MinI(X;Y) \qquad (2)$$

$$C = CommunicationChannelCapacity = MaxI(X;Y) \qquad (3)$$

where D is the distortion which happens when information is compressed (i.e. fewer bits are used to represent or code more frequent or redundant informations) and entropy is the limit to this compression i.e. if one compresses the information beyond the entropy limit there is a high probability that the information will be distorted or erroneous. This is as per Shannon's Source Coding Theorem. We code more frequently used symbols with fewer number of bits and vice-versa.

Mutual information is also the difference of entropies, where entropy is nothing but the measure of uncertainty. Just as entropy (Middleton, 1960) in physical systems tends to increase in the course of time, the reverse is true for information about an information source : as information about the source is processed, it tends to decrease with time, becoming more corrupt or noisy until it is evidently destroyed unless additional information is made available. Here, information refers to the case of desired messages.

3.2 Shannon Entropy

Shannon proposed a measure of uncertainty in a discrete distribution based on the Boltzmann entropy of classical statistical mechanics. He called it the entropy and defined as follows.

We have to take into account the statistics of the alternatives by replacing our original measure of the number of alternatives by the more general expression defining the entropy as follows:

$$H = -\sum_i p_i \log_2(p_i) \qquad (4)$$

where p_i is the probability of the alternative i. The above quantity is known as the binary entropy in *bits* as we use logarithmic base of 2 (with logarithmic base e the entropy is in *nats*), and was shown by Shannon to correspond to the minimum average number of bits needed to encode a probabilistic source of N states distributed with probability p_i. Intuitively, H can also be considered as a measure of uncertainty : it is minimum, and is equal to zero, when one of the alternatives appears with probability one, whereas it is maximum and equals to $\log_2 N$ when all the alternatives are equiprobable so that $p_i = \frac{1}{N}$ for all i.

The term entropy is associated with the uncertainty or randomness whereas information is used to reduce this uncertainty. Uncertainty is the main hindrance to control and if we can reduce the uncertainty by getting the relevant information and utilizing the information properly so as to achieve the desired control performance of the system. Many researchers have posed the same question: *How much information is required for controlling the system based on observed informations in the case where these informations are passed through communication channels in a networked based system?*

Mutual Information $I(X;Y)$ and Entropies $H(X)$, $H(Y)$ and joint entropy $H(X,Y)$ are related as :

$I(X;Y) = H(X) + H(Y) - H(X,Y)$

where $H(X)$ is the uncertainty that X has about Y, $H(Y)$ is the uncertainty that Y has about X , and $H(X,Y)$ is the uncertainty that X and Y hold in common. Information value degrades over time and entropy value increases over time in general. The conditional version of the chain rule (Cover and Thomas, 2006) :

$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$; valid for any random variables X and Y.

Mutual information $I(X;Y)$ is the amount of uncertainty in X, minus the amount of uncertainty in X which remains after Y is known", which is equivalent to "the amount of uncertainty in X which is removed by knowing Y". This corroborates the intuitive meaning of mutual information as the amount of information (that is, reduction in uncertainty) that each variable is having about the other.

The conditional entropy $H(X|Y)$ or read as conditional entropy of X knowing Y or conditioned on Y, is often interpreted in communication theory as representing an information-loss (the so-called equivocation of Shannon (Shannon, 1948)), which results from subtracting the maximum noiseless capacity $I(X;X) = H(X)$ of a communication channel with input X and output Y from the actual capacity of that channel as measured by $I(X;Y)$.

3.3 Bode Integral

Physically an intrinsically stable system needs no information on its internal state or the environment to assure its stability. So, if we consider a well designed stable feedback control system with disturbances or/and noises as inputs and performance signals as outputs then it not needed to have extra feedback loop to assure its stability. We may say the same thing for systems which are intrinsically open-loop stable. For example, a pendulum with non-zero friction coefficient subject to a perturbation will return back to the equilibrium position after a transient period without any need of extra information. For unstable systems the mutual information between the initial state and the output of the system is related to its unstable poles.

The simplest (and perhaps the best known) result is that, for an open loop stable plant, the integral of the logarithm of the closed loop sensitivity is zero; i.e.

$$\int_0^\infty \ln |S_0(j\omega)|\, d\omega = 0$$

Where, S_0 and ω being the sensitivity function and frequency respectively.

Now, we know that the logarithm function has the property that it is negative if $|S_0| < 1$ and it is positive if $|S_0| > 1$. The above result implies that set of frequencies over which sensitivity reduction occurs (i.e. where $|S_0| < 1$) must be matched by a set of frequencies over which sensitivity magnification occurs (i.e. where $|S_0| > 1$). For a stable rational transfer function $L(j\omega)$, sensitivity is defined as $S(j\omega) = \frac{1}{1+L(j\omega)}$. This has been given a nice interpretation as thinking of sensitivity as a pile of dirt. If we remove dirt from one set of frequencies, then it piles up at other frequencies. Hence, if one designs a controller to have low sensitivity in a particular frequency range, then the sensitivity will necessarily increase at other frequencies - a consequence of the weighted integral always being a constant; this phenomenon has also been called the Water-Bed Effect (pushing down on the water bed in one area, raises it somewhere else).

For linear systems Bode Integral is the difference in the entropy rates between the input and output of the systems which is an information-theoretic interpretation. For nonlinear system (if the open loop system is globally exponentially stable and has fading memory) this difference is zero. Fading Memory Requirement is used to limit the contributions of the past values of the input on the output. Entropy of the signals in the feedback loop help provide another interpretation of the Bode integral formula (Zang and Iglesias, 2003)(Mehta et al., 2006) as follows. Shannon En-

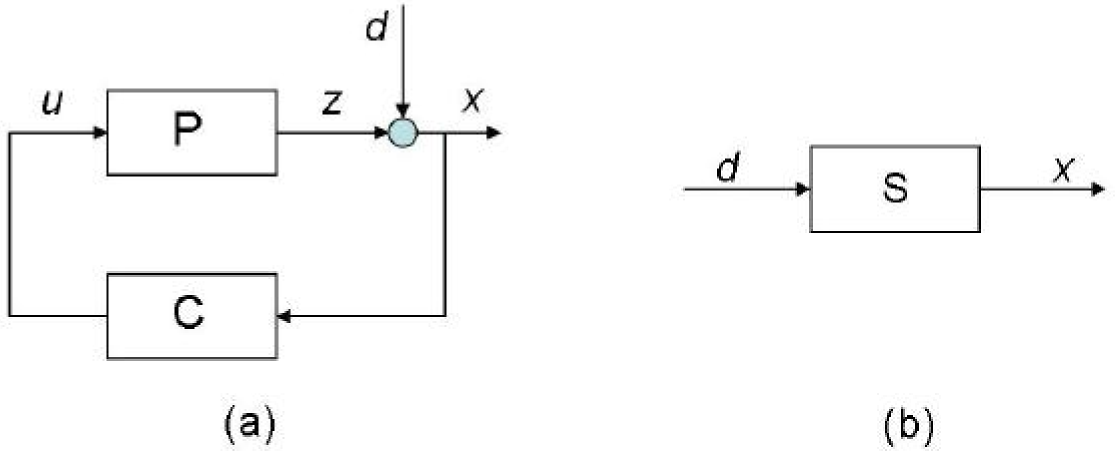


Figure 1: (a) Feedback loop and (b) Sensitivity function.

tropy - Bode Integral Relation can be rewritten as :

$$H_c(x) - H_c(d) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \ln\left|S\left(e^{j\omega}\right)\right| d\omega = \sum_k \log(p_k) \qquad (5)$$

Where $S\left(e^{j\omega}\right)$ is the transfer function of the feedback loop from the disturbance d to output x and p_k 's are unstable poles ($|p_k| > 1$) of the open-loop plant; S is referred to as the sensitivity function for an open-loop plant gain P and a stabilizing feedback controller gain C, S is given by $S = \frac{1}{1+PC}$. Sensitivity shows how much sensitive is the observable output state to input disturbance. Here, $H_c(x)$ and $H_c(d)$ denote the conditional entropy of the random processes associated with the output x and disturbance d respectively as per Figure1(Mehta et al., 2006).

Consider a random variable $x \in \Re^m$ of continuous type with entropy associated with this is given by

$H(x) := -\int_{\Re^m} p(x)\ln p(x)dx;$

where $p(x)$ being the probability density function of x and the conditional entropy of order n is defined as
$H(x_k|x_{k-1},\ldots,x_{k-n}) := -\int_{\Re^m} p(.)\ln p(.)dx$

where $p(.) = (x_k|x_{k-1},\ldots,x_{k-n})$.

This conditional entropy is a measure of the uncertainty about the value of x at time k under the assumption that its n most recent values have been observed. By letting n going to infinity, the conditional entropy of x_k is defined as

$H_c(x_k) := \lim_{n\to\infty} H(x_k|x_{k-1},\ldots,x_{k-n})$ assuming the limit exists. Thus the conditional entropy is a measure of the uncertainty about the value of x at time k under the assumption that its entire past is observed. Difference of conditional entropies between the output and input is nothing but the Bode sensitivity integral which equals the summation of logarithms of unstable poles.

For a stationary Markov process, conditional entropy (Cover and Thomas, 2006) is given by
$H(x_k|x_{k-1},\ldots,x_{k-n}) = H(x_k|x_{k-1})$.

4 RELATED WORK

It is well known that the sensitivity and complementary sensitivity functions represent basic properties of feedback systems such as disturbance attenuation, sensor-noise reduction, and robustness against uncertainties in the plant model. Researchers have worked earlier on the issues of relating the entropy and the Bode Integral and complementary sensitivity. Refer to work in (Sandberg and Bernhardsson, 2005), (Martins et al., 2007), (Bode H., 1945), (Freudenberg and Looze, 1988), (Zang, 2004), (Iglesias, 2002), (Iglesias, 2001), (Zang and Iglesias, 2003), (Mehta et al., 2006), (Sung and Hara, 1989), (Sung and Hara, 1988), (Okano et al., 2008), (Jialing Liu, 2006). In (Iglesias, 2002) the sensitivity integral is interpreted as an entropy integral in the time domain, i.e., no frequency-domain representation is used.

One has to gather relevant information, transmit the information to the relevant agent, process the information, if needed, and then use the information to control the system. The fundamental limitation in information transmission is the achievable information rate (i.e. a fundamental parameter of Information Theory), the fundamental limitation in information processing is the Cramer-Rao Bound (*CRB*) which deals with Fisher Information Matrix (*FIM*) in Estimation Theory, and the fundamental limitation in information utilization is the *Bode Integral* (i.e. a fundamental parameter of Control Theory), seemingly different and usually separately treated, are in fact three sides of the same entity as per the paper (Liu and Elia, 2006). Even Kalman et al. in their paper (Kalman et al., 1963) have stated that Controllability Grammian Matrix **W**-matrix is analogous to *FIM* and the determinant $\det \mathbf{W}$ is analogous to Shannon Information. These research work motivated us to investigate some important correlations amongst mutual information, entropy and design control parameters of practical importance rather than just concentrating on stability issues.

5 INFORMATION INDUCED BY CONTROLLABILITY GRAMMIAN

In general, from the viewpoint of the open-loop system, when the system is unstable, the system amplifies the initial state at a level depending on the size of the unstable poles (Okano et al., 2008). Hence, we can say that in systems having more unstable dynamics, the signals contain more information about the initial state. Using this extra information (in terms of mutual information between the control input and the initial state) we can reduce the entropy (uncertainty) and thus rendering more easy the observation of initial state.
Suppose that we have a feedback control system in which control signal is sent through a network with limited bandwidth. We will consider the case where the state of the system is measurable and the controller can send the state of the system without error. Under these conditions we may write:

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu^*(t); \qquad (6) \\ u^*(t) &= -K_c x(t) + u^e(t); \end{aligned}$$

where K_c, $u^*(t)$ and $u^e(t)$, represent, respectively, the feedback controller gain, the applied control input and control error due to quantization noise of limited bandwidth network. In the sequel we are supposing that the control signal errors are caused by Gaussian White Noise which may be given by $u_i^e(t) = \sqrt{D_i}\delta(t)$. So we may write :

$$\begin{aligned} \dot{x}(t) &= (A - BK_c)x(t) + Bu^e(t); \qquad (7) \\ u^*(t) &= -K_c x(t) + u^e(t); \end{aligned}$$

or more compactly:

$$\begin{aligned} \dot{x}(t) &= A_c x(t) + Bu^e(t); \quad (8) \\ where, A - BK_c = A_c. & \\ u^*(t) &= -K_c x(t) + u^e(t); \end{aligned}$$

The feedback system (8) is a stable one which is perturbed by quantization errors or noises due to the bandwidth limitation.

Lemma : The controllability grammian matrix **W** of system (8) is related with the information-theoretic entropy H as follows (Mitra, 1969):

$$H(x,t) = \frac{1}{2}\ln\{\det \mathbf{W}(\mathbf{D},t)\} + \frac{n}{2}(1 + \ln 2\pi) \qquad (9)$$

(Where **D** being the Diagonal Matrix (positive definite symmetric matrix) with D_i being the ith diagonal element. Here unit impulse inputs are considered.)

= Average *apriori* uncertainty of the state x at time t for an order n of the system.

where

$$\mathbf{W}(\mathbf{D},\tau) = \int_0^\tau e^{A_c t} B \mathbf{D} B^T e^{A_c{}^T t} dt$$

for a system modeled as (8).

Proof of Eqn.(9): Referring to (Cover and Thomas, 2006) we are providing the proof. The input of (8) being Gaussian White Noise, the state of the system is with probability density having mean-value $\bar{x}(t) = e^{A_c t}x(0)$ and Covariance Matrix Σ at time t is given by

$$\Sigma = E\left\{(x - \bar{x})(x - \bar{x})^T\right\} = \mathbf{W}(\mathbf{D},t).$$

In a more detailed form :

$$x(t) = e^{A_c t}x(0) + \int_0^t e^{A_c t} Bu(t)dt$$

$$E\{x(t)\} = E\left\{e^{A_c t}x(0) + \int_0^t e^{A_c t} Bu(t)dt\right\}$$

$$Therefore, \bar{x}(t) = e^{A_c t}x(0)$$

where $\bar{x}(t)$ denotes the mean value of $x(t)$ and Covariance Matrix

$$\Sigma = E\left\{(x - \bar{x})(x - \bar{x})^T\right\}$$

$$\Rightarrow \Sigma = E[\left\{e^{A_c t}x(0) + \int_0^\tau e^{A_c t} Bu(t)dt - e^{A_c t}x(0)\right\}$$
$$\left\{e^{A_c t}x(0) + \int_0^\tau e^{A_c t} Bu(t)dt - e^{A_c t}x(0)\right\}^T]$$

Therefore, $\Sigma = \int_0^\tau e^{A_c t} B\mathbf{D}B^T e^{A_c{}^T t} dt = \mathbf{W}(\mathbf{D},\tau)$.

where, $u_i(t) = \sqrt{D_i}\delta(t)$ i.e. weighted impulses and $\mathbf{D}$ being the Diagonal Matrix (positive definite symmetric matrix) with D_i being the ith diagonal element. Here unit impulse inputs are considered.

$$p(x,t) = \frac{1}{(2\pi)^{n/2}\{\det \mathbf{W}(\mathbf{D},t)\}^{1/2}} e^{[-1/2\left\{(x-\bar{x})^T \mathbf{W}^{-1}(\mathbf{D},t)(x-\bar{x})\right\}]} \tag{10}$$

Now, for multidimensional continuous case, entropy (precisely *differential entropy*) of a continuous random variable X with probability density function $f(x)$ (if $\int_{-\infty}^{\infty} f(x)dx = 1$) is defined (Cover and Thomas, 2006) as
Differential Entropy $h(X) = -\int_S f(x)\ln f(x)dx$;
where the set S for which $f(x) > 0$ is called the *support set* of X.
As in discrete case, the differential entropy depends only on the probability density of the random variable and therefore the differential entropy is sometimes written as $h(f)$ rather than $h(X)$. Here, we call differential entropy as $H(x,t)$ and $f(x)$ as $p(x,t)$ which are correlated as

$$H(x,t) = -\int p(x,t)\ln p(x,t)dx \tag{11}$$

Using equation (10) in equation (11) we get

$$H(x,t) = -\int p(x,t)[-\tfrac{1}{2}(x-\bar{x})^T \mathbf{W}^{-1}(\mathbf{D},t)(x-\bar{x})$$
$$-\ln\left\{(2\pi)^{n/2}\{\det \mathbf{W}(\mathbf{D},t)\}^{1/2}\right\}]dx$$

$$H(x,t) = \tfrac{1}{2}E\left[\sum_{i,j}\{(X_i-\bar{X}_i)(\mathbf{W}^{-1}(\mathbf{D},t))_{ij}(X_j-\bar{X}_j)\}\right]$$
$$+\tfrac{1}{2}\ln[\{(2\pi)^n\{\det \mathbf{W}(\mathbf{D},t)\}]$$

$$= \tfrac{1}{2}E\left[\sum_{i,j}\{(X_i-\bar{X}_i)(X_j-\bar{X}_j)(\mathbf{W}^{-1}(\mathbf{D},t))_{ij}\}\right]$$
$$+\tfrac{1}{2}\ln[\{(2\pi)^n\{\det \mathbf{W}(\mathbf{D},t)\}]$$

$$= \tfrac{1}{2}\sum_{i,j}\left[E\{(X_j-\bar{X}_j)(X_i-\bar{X}_i)\}(\mathbf{W}^{-1}(\mathbf{D},t))_{ij}\right]$$
$$+\tfrac{1}{2}\ln[\{(2\pi)^n\{\det \mathbf{W}(\mathbf{D},t)\}]$$

$$= \tfrac{1}{2}\sum_j\sum_i(\mathbf{W}(\mathbf{D},t))_{ji}(\mathbf{W}^{-1}(\mathbf{D},t))_{ij}$$
$$+\tfrac{1}{2}\ln[\{(2\pi)^n\{\det \mathbf{W}(\mathbf{D},t)\}]$$

$$= \tfrac{1}{2}\sum_j\{(\mathbf{W}(\mathbf{D},t))(\mathbf{W}^{-1}(\mathbf{D},t))\}_{jj}$$
$$+\tfrac{1}{2}\ln[\{(2\pi)^n\{\det \mathbf{W}(\mathbf{D},t)\}]$$

$$= \tfrac{1}{2}\sum_j\mathbf{I}_{jj} + \tfrac{1}{2}\ln[\{(2\pi)^n\}\{\det \mathbf{W}(\mathbf{D},t)\}]$$

(Where $\mathbf{I}_{jj}$ is the Identity Matrix)

$$= \tfrac{n}{2} + \tfrac{1}{2}\ln[\{(2\pi)^n\}\{\det \mathbf{W}(\mathbf{D},t)\}]$$

$$= \tfrac{n}{2} + \tfrac{1}{2}\ln\{(2\pi)^n\} + \tfrac{1}{2}\ln\{\det \mathbf{W}(\mathbf{D},t)\}$$

$$= \tfrac{n}{2} + \tfrac{n}{2}\ln\{(2\pi)\} + \tfrac{1}{2}\ln\{\det \mathbf{W}(\mathbf{D},t)\}$$

$$H(x,t) = \tfrac{1}{2}\ln\{\det \mathbf{W}(\mathbf{D},t)\} + \tfrac{n}{2}(1+\ln 2\pi)$$

Since Controllability Grammian is independent of co-ordinate system and so is the Mutual Information, we try to draw the analogy between the two. Based on the equation (9) we can write the entropy reduction as

$$\Delta H(x,t) = \tfrac{1}{2}\Delta[\ln\{\det \mathbf{W}(\mathbf{D},t)\}]$$

This shows that the entropy reduction which is same as uncertainty reduction is dependent on Controllability Grammian only. Other term being constant for constant n, gets canceled.
Therefore, $\Delta H(x,t) = H(x(t_1),t_1) - H(x(t_2),t_2)$

$$= \tfrac{1}{2}\ln\{\det \mathbf{W}_1(\mathbf{D}_1,t_1)\} - \tfrac{1}{2}\ln\{\det \mathbf{W}_2(\mathbf{D}_2,t_2)\}$$

$$\Rightarrow \Delta H(x,t) = \frac{1}{2}\ln\left\{\frac{\det \mathbf{W}_1(\mathbf{D}_1,t_1)}{\det \mathbf{W}_2(\mathbf{D}_2,t_2)}\right\} \tag{12}$$

For simplicity we denote $\Delta H(x,t)$ by ΔH, $\mathbf{W}_1(\mathbf{D}_1,t_1)$ by $\mathbf{W}_1$ and $\mathbf{W}_2(\mathbf{D}_2,t_2)$ by $\mathbf{W}_2$.

Therefore, $\Delta H = \frac{1}{2}\ln\left\{\det(\frac{\mathbf{W}_1}{\mathbf{W}_2})\right\}$

$\Rightarrow \det(\frac{\mathbf{W}_1}{\mathbf{W}_2}) = e^{2(\Delta H)}$

Using the above expression along with the concept of mutual information being the difference of the entropy and the residual conditional entropy i.e. $I(X;U) = H(X) - H(X|U)$ (gain in information is reduction in entropy), we can conclude that Mutual Information $I(X;U)$ between the state X and control input U denoted simply by Shannon Information I_{sh} is given by this ΔH which can be expressed further as

Finally,

$$\det(\frac{\mathbf{W}_1}{\mathbf{W}_2}) = e^{2(\Delta H)} = e^{2I_{sh}} \tag{13}$$

We may conclude that the uncertainty reduction which is directly related to the $\Delta H(x,t)$ will reduce the variance of the state with respect to the steady-state if $\Delta H(x,t)$ converges to zero. The only influence we have on the control signal is related to that of feedback gain, to be chosen such that the norm of grammians, represented by $\det(W(D_i,t))$ converge rapidly to their norm to infinity $\det(W(D_\infty,\infty))$. We will detail the related approach in a future paper.

6 CONCLUSIONS

This paper has addressed some new ideas concerning the relation between control design and information theory. Since the networked control system has communication constraints due to limited bandwidth or noises, we must have to adopt a policy of resource allocation which enhances the information transmitted. This may be done possible if we know the characteristics of the networks, the bandwidth constraints and that of the dynamical system under study.

As demonstrated the grammian of controllability constitute a metric of information theoretic entropy with respect to the noises induced by quantization. Reduction of these noises is equivalent to the design methods proposing a reduction of the controllability grammian norm. In the case of bandwidth constraints it takes its full interest which will be demonstrated in a future paper. Future work in this direction would be also to propose an information-theoretic analysis for enhancing the zooming algorithm proposed (Ben Gaid and Çela, 2006) and optimal allocation of communication bandwidth which maximizes the systems' performances based on Controllability Grammians. Illustration of these results by simulation and / or experimental verification of the theoretical approaches is the objective of our work.

REFERENCES

Sandberg H. and Bernhardsson B.(2005), A Bode Sensitivity Integral for Linear Time-Periodic Systems, *IEEE Trans. Autom. Control*, Vol. 50, No. 12, December.

Antsaklis P. and Baillieul J. (2007), Special Issue on the Technology of Networked Control Systems, *Proc. of the IEEE*, Vol. 95, No. 1.

Nair G. and Evans R.(2004), Stabilizability of stochastic linear systems with finite feedback data rates, *SIAM J. Control Optim.*, Vol. 43, No. 2, p. 413-436.

Tatikonda S. and Mitter S.(2004), Control under communication constraints, *IEEE Trans. Autom. Control*, Vol. 49, No. 7, p. 1056-1068.

Martins N., Dahleh M., and Doyle J.(2007), Fundamental limitations of disturbance attenuation in the presence of side information, *IEEE Trans. Autom. Control*, Vol. 52, No. 1, p. 56-66.

Bode H.(1945), *Network Analysis and Feedback Amplifier Design*, D. Van Nostrand, 1945.

Iglesias P.(2001), Trade-offs in linear time-varying systems : An analogue of Bode's sensitivity integral, *Automatica*, Vol. 37, No. 10, p. 1541-1550.

Middleton D.(1960), *An Introduction to Statistical Communication Theory*, McGraw-Hill Pub., p. 315.

Cover and Thomas(2006), *Elements of Information Theory*, 2nd Ed., John Wiley and Sons.

Shannon C.(1948), The Mathematical Theory of Communication, *Bell Systems Tech. J.* Vol. 27, p. 379-423, p. 623-656, July, October.

Zang G. and Iglesias P.(2003), Nonlinear extension of Bode's integral based on an information theoretic interpretation, *Systems and Control Letters*, Vol. 50, p. 1129.

Mehta P., Vaidya U. and Banaszuk A.(2006), Markov Chains, Entropy, and Fundamental Limitations in Nonlinear Stabilization, *Proc. of the 45th IEEE Conference on Decision & Control*, San Diego, USA, December 13-15.

Freudenberg J. and Looze D.(1988), Frequency Domain Properties of Scalar and Multivariable Feedback Systems, *Lecture Notes in Control and Information Sciences*, Vol. 104, Springer-Verlag, New York.

Zang G.(2004), Bode's integral extensions in linear time-varying and nonlinear systems, *Ph.D. Dissertation*, Dept. Elect. Comput. Eng., Johns Hopkins Univ., USA.

Iglesias P.(2002), Logarithmic integrals and system dynamics : An analogue of Bode's sensitivity integral for continuous-time time-varying systems, *Linear Alg. Appl.*, Vol. 343-344, p. 451-471.

Sung H. and Hara S.(1989), Properties of complementary sensitivity function in SISO digital control systems, *Int. J. Control*, Vol.50, No. 4, p. 1283-1295.

Sung H. and Hara S.(1988), Properties of sensitivity and complementary sensitivity functions in single-input single-output digital control systems, *Int. J. Control*, Vol. 48, No. 6, p. 2429-2439.

Okano K, Hara S. and Ishii H.(2008), Characterization of a complementary sensitivity property in feedback control: An information theoretic approach, *Proc. of the 17th IFAC-World Congress*, Seoul, Korea, July 6-11, p. 5185-5190.

Jialing L.(2006), Fundamental Limits in Gaussian Channels with Feedback: Confluence of Communication, Estimation and Control, *Ph.D. Thesis*, Iowa State University.

Liu J. and Elia N. (2006), Convergence of Fundamental Limitations in Information, Estimation, and Control, *Proceedings of the 45th IEEE Conference on Decision & Control*, San Diego, USA, December.

Kalman R., Ho Y. and Narendra K.(1963), Controllability of linear dynamical systems, *Contributions to Differential Equations*, Vol. 1, No. 2, p. 189-213.

Mitra D.(1969), W-matrix and the Geometry of Model Equivalence and Reduction, *Proc. of the IEE*, Vol. 116, No.6, June, p. 1101-1106.

Ben Gaid M. M., and Çela A.(2006), Trading Quantization Precision for Sampling Rates in Networked Systems with Limited Communication, *Proceedings of the 45th IEEE Conference on Decision & Control*, San Diego, CA, USA, December 13-15.

OPTIMAL SPARSE CONTROLLER STRUCTURE WITH MINIMUM ROUNDOFF NOISE GAIN

Jinxin Hao, Teck Chew Wee, Lucas S. Karatzas and Yew Fai Lee
School of Engineering, Temasek Polytechnic, 529757, Singapore
jinxin@tp.edu.sg

Keywords: Roundoff noise gain, Sparse controller structure, Optimization, Direct-form II transposed (DFIIt) structure.

Abstract: This paper investigates the roundoff noise effect in the digital controller on the closed-loop output for a discrete-time feedback control system. Based on a polynomial parametrization approach, a sparse controller structure is derived. The performance of the proposed structure is analyzed by deriving the corresponding expression of closed-loop roundoff noise gain and the problem of finding optimized sparse structures is solved. A numerical example is presented to illustrate the design procedure and the performance of the proposed structure compared with those of some existing well-known structures.

1 INTRODUCTION

Finite word length (FWL) effects have been a well studied field in the design of digital filers for more than three decades (Mullis and Roberts, 1976), (Hwang, 1977), (Roberts and Mullis, 1987), (Gevers and Li, 1993). However, they have received less attention in the area of digital control. Nowadays, many researchers have recognized the importance of the numerical problems caused by FWL effects in digital controller implementation. The optimal FWL controller structure design (Fialho and Georgiou, 1994), (Li, 1998), (Wu et al., 2001), (Yu and Ko, 2003) has been considered as one of the most effective methods to minimize the effects of FWL errors on the performance of closed-loop control systems. The basic idea behind this approach is that for a given digital controller, there exist different structures which have different numerical properties, and the optimal structure problem is to identify those structures that optimize a certain FWL performance criterion.

Generally speaking, there are two types of FWL errors in the digital controller. The first one is the perturbation of the controller parameters implemented with FWL, and the second one is the rounding errors that occur in arithmetic operations, which are usually measured with the so-called roundoff noise gain. The effects of roundoff noise have been well studied in digital signal processing, particularly in digital filter implementation (Wong and Ng, 2000), (Wong and Ng, 2001). However, it was not until the late 1980s that the problem of optimal controller realizations minimizing the roundoff noise gain was addressed. The roundoff noise gain was derived for a control system with a state-estimate feedback controller and the corresponding optimal realization problem was solved in (Li and Gevers, 1990), while the roundoff error effect on the linear quadratic regulation (LQG) performance was investigated in (Williamson and Kadiman, 1989) and the optimal solution was obtained by Liu *et al* (Liu et al., 1992). The problem of finding the optimum roundoff noise structures of digital controllers in a sampled-data system has been investigated in (Li et al., 2002).

It has been noted that the optimal controller realizations obtained with the above design methods are usually fully parametrized, which increase the complexity for real-time implementations. From a practical point of view, it is desired that the actually implemented controller have a nice performance against the FWL effects as well as a sparse structure that possesses many trivial parameters[1] which produce no FWL errors. As far as we know, a few results have been published on the sparseness issue for the controller structure design (Li, 1998), (Wu et al., 2003), however, it is noted that in these approaches, sophisticated numerical algorithms were utilized and the positions of trivial parameters were not predictable. In (Hao et al., 2006), we proposed two sparse structures

[1] By trivial parameters we mean those that are 0 and ±1, other parameters are, therefore, referred to as nontrivial parameters.

for digital controllers, which have some degrees of freedom that can be used to enhance the closed-loop stability robustness against the FWL effects.

In this paper, a new sparse controller structure is derived by adopting the polynomial parametrization approach in (Hao et al., 2006) and using the l_2-scaling scheme. This structure can be considered as a l_2-scaled generalized DFIIt (direct-form II transposed) structure. The expression of the roundoff noise gain is derived for a closed-loop feedback control system, in which the digital controller is implemented with the proposed structure. The problem of finding optimized sparse structures is solved by minimizing the corresponding closed-loop roundoff noise gain. A numerical example is given to illustrate the design procedure, which shows that the proposed structure beats the traditional DFIIt structures greatly in terms of roundoff noise performance, and furthermore, outperforms the fully parametrized optimal realization (Li et al., 2002) in terms of both roundoff noise gain and computation efficiency.

2 A SPARSE CONTROLLER STRUCTURE

Consider a discrete-time feedback control system depicted in Fig. 1, where $P_d(z)$ is the discrete-time plant and $C_d(z)$ is a well-designed digital controller. The controller can be represented by its transfer function which is parametrized with $\{\xi_k, \zeta_k\}$ in the shift operator z:

$$C_d(z) = \frac{\sum_{k=0}^{K} \zeta_k z^{K-k}}{z^K + \sum_{k=1}^{K} \xi_k z^{K-k}}. \tag{1}$$

This controller can be implemented with many different structures, such as the direct forms or the following state-space equations:

$$\begin{cases} x(n+1) &= Ax(n) + Bu(n) \\ y(n) &= Cx(n) + du(n) \end{cases} \tag{2}$$

where $x(n) \in \mathcal{R}^{K\times 1}$ is the state variable vector and $u(n)$, $y(n)$ are the input and output of the controller $C_d(z)$, respectively, while $r(n)$ is the input signal of the closed-loop system. $R \triangleq (A, B, C, d)$ is called a realization of $C_d(z)$ with $A \in \mathcal{R}^{K\times K}, B \in \mathcal{R}^{K\times 1}, C \in \mathcal{R}^{1\times K}$ and $d \in \mathcal{R}$, satisfying

$$C_d(z) = d + C(zI - A)^{-1}B.$$

Denote S_C as the set of all the realizations: $S_C \triangleq \{(A, B, C, d) : \ C_d(z) = d + C(zI - A)^{-1}B\}$. Let $R_0 \triangleq (A_0, B_0, C_0, d) \in S_C$ be an initial realization. It can be shown that S_C is characterized by

$$A = T^{-1}A_0T, \ \ B = T^{-1}B_0, \ \ C = C_0T \tag{3}$$

where $T \in \mathcal{R}^{K\times K}$ is any nonsingular matrix. Such a matrix T is usually called a similarity transformation. Once an initial realization R_0 is given, different controller realizations correspond to different similarity transformations T.

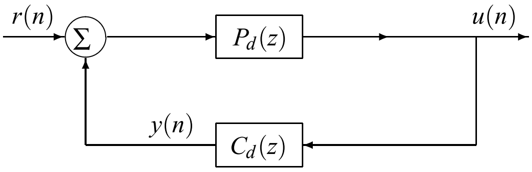


Figure 1: A discrete-time feedback control system.

2.1 A Generalized DFIIt Structure

Based on the approach in (Hao et al., 2006), we define

$$\rho_k(z) \triangleq \frac{z - \gamma_k}{\Delta_k}, \ k = 1, 2, ..., K, \tag{4}$$

where $\{\gamma_k\}$ and $\{\Delta_k \ > 0\}$ are two sets of constants to be discussed later. Let

$$\begin{aligned} p_k(z) &\triangleq \prod_{m=k+1}^{K} \rho_m(z), \ \ \forall k \in \{0, 1, \cdots, K-1\}, \\ p_K(z) &\triangleq 1. \end{aligned} \tag{5}$$

It can be shown that (1) can be rewritten as

$$C_d(z) = \frac{\beta_0 p_0(z) + \beta_1 p_1(z) + \ldots + \beta_K p_K(z)}{p_0(z) + \alpha_1 p_1(z) + \ldots + \alpha_K p_K(z)}, \tag{6}$$

where

$$\begin{aligned} \bar{\alpha} &\triangleq [1 \quad \alpha_1 \quad \cdots \quad \alpha_K]^{\mathcal{T}} \\ &= \kappa \bar{T}_p^{-\mathcal{T}} [1 \quad \xi_1 \quad \cdots \quad \xi_K]^{\mathcal{T}} \\ \bar{\beta} &\triangleq [\beta_0 \quad \beta_1 \quad \cdots \quad \beta_K]^{\mathcal{T}} \\ &= \kappa \bar{T}_p^{-\mathcal{T}} [\zeta_0 \quad \zeta_1 \quad \cdots \quad \zeta_K]^{\mathcal{T}} \end{aligned}$$

with $\kappa = \prod_{k=1}^{K} \Delta_k^{-1}$ such that $\bar{\alpha}(1) = 1$ and $\bar{T}_p$ an upper triangular matrix whose kth row is formed with the coefficients of $p_{k-1}(z)$ defined above. Equation (6) implies that the controller transfer function $C_d(z)$ is reparametrized with $\{\alpha_k\}$ and $\{\beta_k\}$ in the new set of polynomial operators $\{p_k(z)\}$.

It follows from (5) and (6) that the output of the controller can be computed with the following equations

$$\begin{aligned} y(n) &= \beta_0 u(n) + w_1(n) \\ w_k(n) &= \rho_k^{-1}[\beta_k u(n) - \alpha_k y(n) + w_{k+1}(n)] \\ w_K(n) &= \rho_K^{-1}[\beta_K u(n) - \alpha_K y(n)] \end{aligned} \tag{7}$$

where $w_k(n)$ is the output of $\rho_k^{-1}(z)$ and can be computed with the structure depicted in Fig. 2. Fig. 3 shows the corresponding structure to (7). For convenience, a structure defined by Fig.s 2 and 3 is called a generalized DFIIt structure, denoted as ρDFIIt. This structure possesses $\{\alpha_k,\ \beta_k,\ \Delta_k\}$ and a set of free parameters $\{\gamma_k\}$. For a given digital controller $C_d(z)$, there exists a class of such structures, depending on the space within which $\{\gamma_k\}$ take values. Clearly, when $\gamma_k = 0,\ \Delta_k = 1,\ \forall k$, Fig. 3 is the conventional direct-form II transposed (DFIIt) structure.

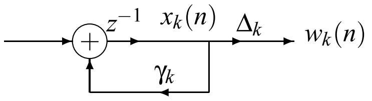


Figure 2: A realization of $\rho_k^{-1}(z)$ defined in (4).

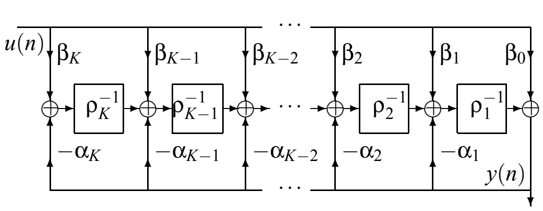


Figure 3: Block diagram of the ρDFIIt structure.

With $\{x_k(n)\}$ indicated in Fig. 2 as the state variables and $x(n)$ denoting the state vector, one can find the equivalent state-space realization, denoted as $(A_\rho, B_\rho, C_\rho, \beta_0)$, of the proposed ρDFIIt structure:

$$C_d(z) = \beta_0 + C_\rho(zI - A_\rho)^{-1}B_\rho \tag{8}$$

with $B_\rho = \bar{V}_\beta - \beta_0\bar{V}_\alpha$, where $\bar{V}_x \triangleq [x_1\ \cdots\ x_k\ \cdots\ x_K]^{\mathcal{T}}$ for $x = \alpha, \beta$, $C_\rho = [\Delta_1\ 0\ \cdots\ 0\ 0]$, and

$$A_\rho \triangleq \begin{bmatrix} a_{11} & \Delta_2 & 0 & \cdots & 0 & 0 \\ a_{21} & \gamma_2 & \Delta_3 & \cdots & 0 & 0 \\ & & & \vdots & & \\ a_{(K-1)1} & 0 & 0 & \cdots & \gamma_{K-1} & \Delta_K \\ a_{K1} & 0 & 0 & \cdots & 0 & \gamma_K \end{bmatrix}$$

with $a_{11} = \gamma_1 - \Delta_1\alpha_1$ and $a_{k1} = -\Delta_1\alpha_k,\ k \in \{2, 3, \cdots, K\}$.

2.2 Scaling Scheme

It is well known that in an implementation system, all the signals should be sustained within a certain dynamic range in order to avoid overflow. Under the assumption that the input $r(n)$ and the output $u(n)$ of the closed-loop system are properly pre-scaled, the only signals which may have overflow are the elements of the controller state vector $x(n)$, which, therefore, have to be scaled.

There exist different scaling schemes for preventing variables from overflow. The popularly used ones are the l_2- and l_∞-scalings. In what follows, we will concentrate on the l_2-scaling scheme. The l_2-scaling means that each element of the controller state vector $x(n)$ should have a unit variance when the input $r(n)$ is a white noise with a unit variance. This can be achieved if

$$\bar{\mathcal{K}}(l,l) = 1,\ \ l = N+1, N+2, ..., N+K \tag{9}$$

where $\bar{\mathcal{K}}$ is the controllability Gramian of the closed-loop system of order $N+K$. Assuming that $P_d(z)$ is strictly proper and has a realization $(A_z, B_z, C_z, 0)$, let $(A_{cl}, B_{cl}, C_{cl}, 0)$ be the closed-loop realization, where

$$\begin{aligned} A_{cl} &= \begin{bmatrix} A_z + dB_zC_z & B_zC \\ BC_z & A \end{bmatrix} \\ B_{cl} &= \begin{bmatrix} B_z \\ \mathbf{0} \end{bmatrix} \\ C_{cl} &= [C_z \quad \mathbf{0}] \end{aligned} \tag{10}$$

with $\mathbf{0}$ denoting the zero vector of appropriate dimension. Then $\bar{\mathcal{K}}$ is given by

$$\bar{\mathcal{K}} = \sum_{k=0}^{+\infty} A_{cl}^k B_{cl} B_{cl}^{\mathcal{T}} (A_{cl}^{\mathcal{T}})^k \tag{11}$$

satisfying

$$\bar{\mathcal{K}} = A_{cl}\bar{\mathcal{K}}A_{cl}^{\mathcal{T}} + B_{cl}B_{cl}^{\mathcal{T}}.$$

Let (A_{cl}, B_{cl}, C_{cl}) and $(A_{cl}^0, B_{cl}^0, C_{cl}^0)$ be two realizations of the closed-loop system with A_{cl}, B_{cl} and C_{cl} defined in (10), corresponding to the two digital controller realizations $R \triangleq (A, B, C, d)$ and $R_0 \triangleq (A_0, B_0, C_0, d)$ which are related with (3), respectively. It can be shown that

$$\begin{aligned} A_{cl} &= \begin{bmatrix} I & \mathbf{0} \\ \mathbf{0} & T \end{bmatrix}^{-1} A_{cl}^0 \begin{bmatrix} I & \mathbf{0} \\ \mathbf{0} & T \end{bmatrix} \\ B_{cl} &= \begin{bmatrix} I & \mathbf{0} \\ \mathbf{0} & T \end{bmatrix}^{-1} B_{cl}^0 \\ C_{cl} &= C_{cl}^0 \begin{bmatrix} I & \mathbf{0} \\ \mathbf{0} & T \end{bmatrix}. \end{aligned} \tag{12}$$

It then follows from (12) that

$$\bar{\mathcal{K}} = \begin{bmatrix} I & \mathbf{0} \\ \mathbf{0} & T \end{bmatrix}^{-1} \bar{\mathcal{K}}_0 \begin{bmatrix} I & \mathbf{0} \\ \mathbf{0} & T \end{bmatrix}^{-\mathcal{T}}$$

where $\bar{\mathcal{K}}^0$ is the closed-loop controllability Gramian corresponding to R_0. Let

$$\bar{\mathcal{K}} \triangleq \begin{bmatrix} \mathcal{K}_{11} & \mathcal{K}_{12} \\ \mathcal{K}_{21} & \mathcal{K} \end{bmatrix},\ \ \bar{\mathcal{K}}^0 \triangleq \begin{bmatrix} \mathcal{K}_{11}^0 & \mathcal{K}_{12}^0 \\ \mathcal{K}_{21}^0 & \mathcal{K}_0 \end{bmatrix} \tag{13}$$

have the same partition as $\begin{bmatrix} I & \mathbf{0} \\ \mathbf{0} & T \end{bmatrix}$, then

$$\mathcal{K} = T^{-1}\mathcal{K}_0 T^{-\mathcal{T}} \qquad (14)$$

where $\mathcal{K}_0$ is a positive-definite matrix independent of T.

It is easy to see from above equations that the l_2-scaling constraint (9) can be satisfied if the diagonal elements of $\mathcal{K}$ are all equal to one, that is

$$\mathcal{K}(k,k) = 1, \forall k. \qquad (15)$$

When the ρDFIIt structure is used to implement a digital controller, it has to be l_2-scaled in order to prevent the signals in the controller from overflow, which can be achieved by choosing $\{\Delta_k\}$ properly. It is interesting to note that

$$p_k(z) = [\prod_{l=k+1}^{K} \Delta_l^{-1}]\bar{p}_k(z),\ \forall k \qquad (16)$$

where all $\bar{p}_k(z)$ are obtained using (5) with $\Delta_k = 1,\ \forall k$.

Let $(A_\rho^0, B_\rho^0, C_\rho^0, \beta_0)$ be the equivalent state-space realization corresponding to $\Delta_k = 1,\ \forall k$. With (16), it can be shown that

$$A_\rho = T_{sc}A_\rho^0 T_{sc}^{-1},\ B_\rho = T_{sc}B_\rho^0,\ C_\rho = C_\rho^0 T_{sc}^{-1}$$

where T_{sc}^{-1} is a diagonal scaling similarity transformation, and

$$T_{sc} = diag(d_1, d_2, \cdots, d_K),\ \ d_k = \prod_{l=1}^{k} \Delta_l^{-1},\ \forall k.$$

Denote $\bar{\mathcal{K}}_\rho$ and $\bar{\mathcal{K}}_\rho^0$ as the closed-loop controllability Gramians, corresponding to the controller realizations $(A_\rho, B_\rho, C_\rho, \beta_0)$ and $(A_\rho^0, B_\rho^0, C_\rho^0, \underline{\beta}_0)$, respectively. Let $\mathcal{K}_\rho$ be the sub-matrix of $\bar{\mathcal{K}}_\rho$ with the partition defined in (13), then (14) becomes $\mathcal{K}_\rho = \underline{T_{sc}}\mathcal{K}_\rho^0 T_{sc}^{\mathcal{T}}$ with $\mathcal{K}_\rho^0$ the corresponding sub-matrix of $\bar{\mathcal{K}}_\rho^0$. It is easy to see that the l_2-scaling can be achieved if $\mathcal{K}_\rho(k,k) = 1, \forall k$, or equivalently,

$$d_k^2 \mathcal{K}_\rho^0(k,k) = 1,\ \ k = 1, 2, ..., K$$

which leads to

$$\Delta_1 = \sqrt{\mathcal{K}_\rho^0(1,1)},\ \Delta_k = \sqrt{\frac{\mathcal{K}_\rho^0(k,k)}{\mathcal{K}_\rho^0(k-1,k-1)}}, \qquad (17)$$

$$k = 2, 3, ..., K.$$

In the sequel, all the structures under discussion, including the ρDFIIt structure, are assumed to have been l_2-scaled. Here we should note that the l_2-scaled ρDFIIt structure to be analyzed in this paper is different from the structure in (Hao et al., 2006) where $\{\Delta_k\}$ are free parameters used for maximizing the stability robustness measure.

3 PERFORMANCE ANALYSIS AND OPTIMIZED STRUCTURE

In this section, we will analyze the performance of the ρDFIIt structure in terms of closed-loop roundoff noise gain. The problem of finding the optimized structure will then be formulated and solved.

One notes that for a given digital controller $C_d(z)$, there exists a class of l_2-scaled ρDFIIt structures, which are determined by a space, denoted as S_γ, from which the free parameters $\{\gamma_k\}$ take values. It is easy to see that $\{\gamma_k\}$ are the parameters to be implemented directly in the structure. Since we are confined to fixed-point implementation for which the FWL effects are more serious, it is desired that γ_k be absolutely not bigger than one and of FWL format. For a fixed-point implementation of B_p bits, define

$$S_{FWL} \triangleq \{-1, 1\} \cup \{\pm \sum_{l=1}^{B_p} b_l 2^{-l},\ b_l = 0, 1,\ \forall l\} \quad (18)$$

which is a discrete space, containing $2^{B_p+1}+1$ elements. Therefore, one can choose $S_\gamma \subset S_{FWL}$, which means that all γ_k are of exact B_γ-bit format with $B_\gamma \leq B_p$.

3.1 Closed-loop Roundoff Noise Gain

In practice, a designed digital controller has to be implemented with finite precision and a rounding operation has to be applied if less-than-double precision fixed-point arithmetic is utilized. Assuming rounding occurs after multiplication (RAM), a variable, say x, computed with a multiplication, has to be replaced by its quantized version, denoted as $Q[x]$, in the ideal computation model. The difference $Q[x] - x$ is the corresponding roundoff noise, which is usually modelled as a white noise sequence and statistically independent of those produced by other sources.

Let μ be a parameter in a controller structure and $Q[\mu s(n)]$ the quantized version of the product $\mu s(n)$. The roundoff noise due to the parameter μ can be defined as

$$\psi(\mu)\varepsilon_\mu(n) \triangleq Q[\mu s(n)] - \mu s(n)$$

where $\psi(\mu) = 1$ if μ is nontrivial, otherwise, $\psi(\mu) = 0$. In fact, the function $\psi(\mu)$ is used for indicating the fact that μ produces no roundoff noise when it is trivial. Denote $\Delta u(n)$ as the corresponding output deviation of the closed-loop system to $\psi(\mu)\varepsilon_\mu(n)$ and $F(z)$ as the transfer function between $\psi(\mu)\varepsilon_\mu(n)$ and $\Delta u(n)$. It is well known (see, e.g., (Gevers and Li, 1993)) that $\Delta u(n)$ is a stationary process and the variance $E[(\Delta u(n))^2] = \psi(\mu)\|F(z)\|_2^2 E[\varepsilon_\mu^2(n)]$. Then the roundoff noise gain for μ is defined as

$$G_\mu \triangleq \frac{E[(\Delta u(n))^2]}{E[\varepsilon_\mu^2(n)]} = \psi(\mu)||F(z)||_2^2 \qquad (19)$$

where $||.||_2$ is the L_2-norm:

$$\begin{aligned} \| F(z) \|_2 &\triangleq \left\{ \frac{1}{2\pi}\int_0^{2\pi} \sum_{i=1}^{l}\sum_{k=1}^{m} |f_{ik}(e^{j\omega})|^2 d\omega \right\}^{1/2} \\ &= \left\{ tr[\frac{1}{j2\pi}\oint_{|z|=1} F(z)F^{\mathcal{H}}(z)z^{-1}dz] \right\}^{1/2} \quad (20) \end{aligned}$$

with $F(z) = \{f_{ik}(z)\} \in \mathcal{R}^{l\times m}$, and $\mathcal{H}$, $tr[.]$ denoting the conjugate-transpose and trace operators, respectively. Let $F(z) = D + L(zI-\Phi)^{-1}J$. It can be shown that

$$||F(z)||_2^2 = tr[DD^{\mathcal{T}} + LW_cL^{\mathcal{T}}] = tr[D^{\mathcal{T}}D + J^{\mathcal{T}}W_oJ] \qquad (21)$$

where W_c, W_o are the controllability and observability Gramians of the realization (Φ, J, L, D), respectively.

Consider a digital controller implemented with a ρDFIIt structure. We note that the parameters in a ρDFIIt structure are $\{\alpha_k\}, \{\beta_k\}$, $\{\Delta_k\}$, and $\{\gamma_k\}$. It follows from (6) that

$$y(n) = \beta_0 u(n) + \sum_{l=1}^{K}[\frac{p_l(z)}{p_0(z)}\beta_l u(n) - \frac{p_l(z)}{p_0(z)}\alpha_l y(n)]. \qquad (22)$$

Let us first look at the effect of roundoff noise $\psi(\beta_0)\varepsilon_{\beta_0}(n)$ due to β_0 on the closed-loop output. Let $u^*(n)$ and $y^*(n)$ be the corresponding output of the closed-loop system and the controller, respectively. Clearly, they obey (22) with $\beta_0 u^*(n)$ replaced by $\beta_0 u^*(n) + \psi(\beta_0)\varepsilon_{\beta_0}(n)$. Denote $\Delta y(n) \triangleq y^*(n) - y(n)$. Then one can show that

$$\begin{aligned} \Delta y(n) = [\beta_0\Delta u(n) + \psi(\beta_0)\varepsilon_{\beta_0}(n)] \\ + \sum_{l=1}^{K}\frac{p_l(z)}{p_0(z)}\beta_l\Delta u(n) - \sum_{l=1}^{K}\frac{p_l(z)}{p_0(z)}\alpha_l\Delta y(n) \qquad (23) \end{aligned}$$

where $\Delta u(n) \triangleq u^*(n) - u(n)$, satisfying

$$\Delta u(n) = P_d(z)\Delta y(n). \qquad (24)$$

Let $H_{cl}(z)$ be the transfer function of the closed-loop system, which is given by

$$H_{cl}(z) = \frac{P_d(z)}{1 - P_d(z)C_d(z)}$$

where $P_d(z)$ is the transfer function of plant and $C_d(z)$ the polynomial parametrized controller transfer function given by (6). It is easy to see that

$$H_{cl}(z) = D_{cl} + C_{cl}(zI - A_{cl})^{-1}B_{cl} \qquad (25)$$

with $(A_{cl}, B_{cl}, C_{cl}, D_{cl})$ the realization of closed-loop system. It then follows from (23) and (24) that

$$\Delta u(n) = S_0(z)\psi(\beta_0)\varepsilon_{\beta_0}(n)$$

where $S_0(z)$ is the transfer function between $\psi(\beta_0)\varepsilon_{\beta_0}(n)$ and $\Delta u(n)$, which is given by

$$S_0(z) = H_{cl}(z)V_0(z)$$

with

$$V_0(z) \triangleq \frac{p_0(z)}{p_0(z) + \sum_{l=1}^{K}\alpha_l p_l(z)}.$$

Comparing $V_0(z)$ with (6), it follows from (8) that

$$\begin{aligned} V_0(z) &= [\beta_0 + C_\rho(zI - A_\rho)^{-1}B_\rho]_{|\beta_0=1,\bar{V}_\beta=\mathbf{0}} \\ &= 1 - C_\rho(zI - A_\rho)^{-1}\bar{V}_\alpha. \end{aligned}$$

One observes that $S_0(z)$ is of the form $S_0(z) = [D_2 + C_2(zI_2 - A_2)^{-1}B_2][D_1 + C_1(zI_1 - A_1)^{-1}B_1]$, where $A_1 = A_\rho, B_1 = -V_\alpha, C_1 = C_\rho, D_1 = 1$, $A_2 = A_{cl}, B_2 = B_{cl}, C_2 = C_{cl}, D_2 = D_{cl}$, and $I_k, k = 1,2$ denotes the identity matrix of a proper dimension. It is easy to verify that

$$S_0(z) \triangleq \tilde{D} + \tilde{C}(z\tilde{I} - \tilde{A})^{-1}\tilde{B}$$

where

$$\begin{aligned} \tilde{D} &= D_2D_1,\ \tilde{C} = [D_2C_1 \quad C_2] \\ \tilde{I} &= \begin{bmatrix} I_1 & \mathbf{0} \\ \mathbf{0} & I_2 \end{bmatrix} \\ \tilde{A} &= \begin{bmatrix} A_1 & \mathbf{0} \\ B_2C_1 & A_2 \end{bmatrix},\ \tilde{B} = \begin{bmatrix} B_1 \\ B_2D_1 \end{bmatrix}. \end{aligned}$$

According to (19) and (21), the roundoff noise gain due to parameter β_0 is given by

$$\begin{aligned} G_{\beta_0} &= \psi(\beta_0)||S_0(z)||_2^2 = \psi(\beta_0)tr(\tilde{D}^{\mathcal{T}}\tilde{D} + \tilde{B}^{\mathcal{T}}\tilde{W}\tilde{B}) \\ &\triangleq \psi(\beta_0)G_0 \end{aligned}$$

where $\tilde{W}$ is the observability Gramian of the realization $(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D})$.

Using the same procedure, one can analyze the roundoff noise gain due to the parameter β_k. Let $\psi(\beta_k)\varepsilon_{\beta_k}(n)$ be the corresponding roundoff noise. It can be shown that the transfer function from $\psi(\beta_k)\varepsilon_{\beta_k}(n)$ to $\Delta u(n)$, denoted as $S_k(z)$, is

$$S_k(z) = H_{cl}(z)V_k(z)$$

with $H_{cl}(z)$ given by (25) and

$$V_k(z) = \frac{p_k(z)}{p_0(z) + \sum_{l=1}^{K}\alpha_l p_l(z)} = C_\rho(zI - A_\rho)^{-1}e_k$$

for $k = 1, 2, \cdots, K$, where e_k is the kth elementary vector whose elements are all zero except the kth one which is 1. Therefore,

$$G_{\beta_k} = \psi(\beta_k)||S_k(z)||_2^2 \triangleq \psi(\beta_k)G_k,\ \forall k$$

with

$$G_k = tr(\tilde{D}_k^T \tilde{D}_k + \tilde{B}_k^T \tilde{W}_k \tilde{B}_k)$$

where $(\tilde{A}_k, \tilde{B}_k, \tilde{C}_k, \tilde{D}_k)$ is the realization of $S_k(z)$ and $\tilde{W}_k$ is the corresponding observability Gramian.

Comparing the positions of α_k, γ_k and Δ_{k+1} with that of β_k in Fig. 3, one can see easily that

$$G_{\alpha_k} = \psi(\alpha_k)G_k, \; G_{\gamma_k} = \psi(\gamma_k)G_k, \; G_{\Delta_k} = \psi(\Delta_k)G_{k-1}$$

for $k = 1, 2, \cdots, K$.

Therefore, the total closed-loop roundoff noise gain of the ρDFIIt structure is

$$G_\rho \triangleq \sum_{k=1}^{K} [G_{\alpha_k} + G_{\gamma_k} + G_{\Delta_k}] + \sum_{k=0}^{K} G_{\beta_k} \triangleq \sum_{k=0}^{K} \upsilon_k G_k \quad (26)$$

where the coefficients υ_k can be specified easily with the expressions, obtained above, of roundoff noise gain for all the parameters.

3.2 Structure Optimization

For a given digital controller $C_d(z)$ and any given free parameters $\{\gamma_k\}$, one can obtain the l_2-scaled ρDFIIt structure with the procedure presented in Section II. The roundoff noise gain G_ρ can then be evaluated with (26). Since different sets of $\{\gamma_k\}$ yield different ρDFIIt structures and hence lead to different roundoff noise gain G_ρ, an interesting problem is to minimize G_ρ with respect to these free parameters, which leads to the following optimal ρDFIIt structure problem:

$$\min_{\gamma_k \in S_\gamma} G_\rho. \quad (27)$$

It seems impossible to obtain analytical solutions to the problem (27) due to the high nonlinearity of G_ρ in $\{\gamma_k\}$. However, noting that S_γ is of finite number of elements, the problem can be well solved using the exhaustive searching method.

4 A DESIGN EXAMPLE

In this section, we illustrate our design procedure and the performance of the proposed structure with a numerical example, in which $S_\gamma = \{\pm 1, \pm(2^{-1} + 2^{-2}), \pm 2^{-1}, \pm 2^{-2}, 0\}$. The elements in the set S_γ are of exact 3-bit fixed-point format (including one bit for the sign). Using more bits or floating-point formats will lead to a further improved performance, which can also confirm the effectiveness of our design procedure.

Consider a discrete-time control system, where the digital plant $P_d(z) = 10^{-1} \times \frac{0.0181z^4+0.0033z^3-0.1628z^2+0.0111z+0.0163}{z^5-3.7174z^4+5.7458z^3-4.6673z^2+2.0336z-0.3953}$ and controller $C_d(z) = 0.0577 + \frac{0.2258z^5-0.6588z^4+0.8195z^3-0.5320z^2+0.1814z-0.0234}{z^6-3.6172z^5+5.9513z^4-5.6335z^3+3.2509z^2-1.0895z+0.1690}$.

Table 1: Comparison of Different Structures.

	zDFIIt	δDFIIt	R_f	ρDFIIt
G	1.5191×10^4	7.1763	4.9919	1.0085
N_p	19	19	49	24

The corresponding poles of the closed-loop system are $\{0.4523 \pm j0.5315, 0.4837 \pm j0.4556, 0.6055 \pm j0.4108, 0.7814 \pm j0.3099, 0.8886 \pm j0.3326, 0.9113\}$.

Applying exhaustive searching to (27), one gets the optimal ρDFIIt structure, denoted as ρDFIIt, for which $\gamma_1 = 1, \; \gamma_5 = 0.5, \; \gamma_k = 0.75, \; k \in \{2, 3, 4, 6\}$. For comparison, an optimal fully parametrized state-space realization, denoted by R_f, is obtained using the procedure in (Li et al., 2002). zDFIIt and δDFIIt are the traditional DFIIt structures in the shift- and δ-operators, corresponding to $\gamma_k = 0, \; \forall k$ and $\gamma_k = 1, \; \forall k$, respectively.

The comparative results of different structures are presented in Table I, where G is the roundoff noise gain and N_p is the number of nontrivial parameters in each structure.

From this example, one can see that zDFIIt yields a very large roundoff noise gain, though it has only 19 parameters to implement, while δDFIIt has a much better performance. The fully parametrized optimal realization R_f yields a further better performance, however, all the 49 parameters in R_f are nontrivial. It is interesting to see that ρDFIIt beats R_f in terms of the roundoff noise performance. Moreover, ρDFIIt is very sparse and has only 24 nontrivial parameters, which is less than half of those in R_f.

5 CONCLUSIONS

In this paper, we have addressed the optimal controller structure problem in a discrete-time control system with roundoff noise consideration. Our major contribution is twofold. Firstly, a sparse controller structure, which is a l_2-scaled generalized DFIIt structure, has been derived. Secondly, the performance of the proposed structure has been analyzed by deriving the corresponding expression of closed-loop roundoff noise gain and the problem of finding optimized sparse structures has been solved. Finally, a numerical example has been given, which shows that the proposed structure can achieve much better performance than some well-known structures and particularly, outperforms the traditional optimal fully parametrized realization greatly in terms of reducing roundoff noise and implementation complexity. This optimal controller design strategy with high precision arithmetic can be utilized to develop suitable control systems for robotic platforms performing complex movements, where efficiency, accuracy and fast speed are essential.

REFERENCES

Fialho, I. J. and Georgiou, T. T. (1994). On stability and performance of sampled-data systems subject to wordlength constraint. *IEEE Trans. Automat. Contr.*, 39:2476–2481.

Gevers, M. and Li, G. (1993). *Parametrizations in Control, Estimation and Filtering Problems: Accuracy Aspects*. Springer-Verlag, London, U.K.

Hao, J., Li, G., and Wan, C. (2006). Two classes of efficient digital controller structures with stability consideration. *IEEE Trans. Automat. Contr.*, 51:164–170.

Hwang, S. Y. (1977). Minimum uncorrelated unit noise in state-space digital filtering. *IEEE Trans. Acoust., Speech, Signal Processing*, 25:273–281.

Li, G. (1998). On the structure of digital controllers with finite word length consideration. *IEEE Trans. Automat. Contr.*, 43:689–693.

Li, G. and Gevers, M. (1990). Optimal finite-precision implementation of a state-estimate feedback controller. *IEEE Trans. Circuits Syst.*, 38:1487–1499.

Li, G., Wu, J., Chen, S., and Zhao, K. Y. (2002). Optimum structures of digital controllers in sampled-data systems: a roundoff noise analysis. *IEE Proc. Control Theory Appl.*, 149:247–255.

Liu, K., Skelton, R., and Grigoriadis, K. (1992). Optimal controllers for finite wordlength implementation. *IEEE Trans. Automat. Contr.*, 37:1294–1304.

Mullis, C. T. and Roberts, R. A. (1976). Synthesis of minimum roundoff noise fixed-point digital filters. *IEEE Trans. Circuits Syst.*, 23:551–562.

Roberts, R. A. and Mullis, C. T. (1987). *Digital Signal Processing*. Addison Wesley.

Williamson, D. and Kadiman, K. (1989). Optimal finite wordlength linear quadratic regulation. *IEEE Trans. Automat. Contr.*, 34:1218–1228.

Wong, N. and Ng, T. S. (2000). Roundoff noise minimization in a modified direct form delta operator iir structure. *IEEE Trans. Circuits Syst. II*, 47:1533–1536.

Wong, N. and Ng, T. S. (2001). A generalized direct-form delta operator-based iir filter with minimum noise gain and sensitivity. *IEEE Trans. Circuit Syst. II*, 48:425–431.

Wu, J., Chen, S., Li, G., and Chu, J. (2003). Constructing sparse realisations of finite-precision digital controllers based on a closed-loop stability related measure. *IEE Proc. Control Theory Appl.*, 150:61–68.

Wu, J., Chen, S., Li, G., Istepanian, R. H., and Chu, J. (2001). An improved closed-loop stability related measure for finite-precision digital controller realizations. *IEEE Trans. Automat. Contr.*, 46:1162–1166.

Yu, W. S. and Ko, H. J. (2003). Improved eigenvalue sensitivity for finite-precision digital controller realizations via orthogonal hermitian transform. *IEE Proc. Control Theory Appl.*, 50:365–375.

ON THE STATE–SPACE REALIZATION OF VECTOR AUTOREGRESSIVE STRUCTURES
An Assessment

Vasilis K. Dertimanis, Dimitris V. Koulocheris
Vehicles Laboratory, National Technical University of Athens, Iroon Politechniou 9, 157 80, Athens, Greece
bullit@central.ntua.gr, dbkoulva@central.ntua.gr

Keywords: Vector autoregressive, Time–series, State–space, Green function, Covariance matrix, Dispersion analysis, Estimation.

Abstract: This study explores the interconnection between vector autoregressive (VAR) structures and state–space models and results in a compact framework for the representation of multivariate time–series, as well as the estimation of structural information. The corresponding methodology that is developed, applies the fact that every VAR process of order *n* may be described by an equivalent (non–unique) VAR model of first order, which is identical to a state–space realization. The latter uncovers many ”hidden” information of the initial model, it is more easy to manipulate and maintains significant second moments’ information that can be reflected back to the original structure with no effort. The performance of the proposed framework is validated using vector time–series signatures from a structural system with two degrees of freedom, which retains a pair of closely spaced vibration modes and has been reported in the relevant literature.

1 INTRODUCTION

The analysis of vector time–series, generally referring to the determination of the dynamics that govern the performance of a system under unobservable excitations, has been a subject of constant development for more than two decades, as part of the broader system identification framework. Relative applications are extended from econometrics (Clements and Henry, 1998; Lütkepohl, 2005), to dynamics (Ljung, 1999; Koulocheris et al., 2008), vibration (Papakos and Fassois, 2003), modal analysis (Huang, 2001) and fault diagnosis (Dertimanis, 2006).

The study of vector time–series can be assessed from a variety of viewpoints, with respect to the application of interest. These include simulation, prediction and extraction of structural information. Yet, while in the first two areas the interrelation of the corresponding time–series structures, such as the VAR one (or the VARX and the VARMAX, under the availability of input information), to equivalent state–space models has been studied extensively (Hannan, 1976; Brockwell and Davis, 2002; Lütkepohl, 2005), not much have been done in the third (Lardies, 2008), from where it appears that state–space realizations may provide significant advantages, regarding structural estimation, with respect to other approaches (He and Fu, 2001).

This paper attempts to provide a unified framework for the representation of vector time–series, by means of VAR structures and their corresponding state–space realizations. Based on the fact that every VAR structure of order *n* (referred to from now on as VAR(n) structure) can be expressed as an equivalent (and non–unique) VAR(1) one, a corresponding state–space model is developed. This specific model qualifies, over other possible realizations, for having a transition matrix that coincides with the VAR(n) polynomial matrix. It turns out that the spectrum of this transition matrix has all the structural information about the system that generates the time–series ”hidden” in its spectrum. Consequently, by taking advantage standard results of matrix algebra, closed form expressions for the Green function and the covariance matrix are derived. The latter, unlike other estimation schemes, such as the Burg and the forward–backward methods (Brockwell and Davis, 2002), is by definition closely related to the energy distribution of the vector time–series. Thus, the corresponding expression that is assessed, quantifies the impact of each specific structural mode in the total energy of the system, a technique that has been recorded in the literature as *dispersion analysis* (Lee and Fassois, 1993).

The paper is organized as follows: in Sec. 2 the VAR(n) structure is presented and the reduction to the state–space realization is performed. Section 3 illustrates the properties of the state–space model, including the development of closed form representations for the Green function and the covariance matrix, and how these are reflected to the original VAR(n) structure. Section 4 contains the least–squares estimation of the state equation and Sec. 5 the validation of the estimated model, as well as the extraction of the structural information that is ”hidden” in the transition matrix. Section 6 displays an application of the proposed framework to a simulated vibrating system that has been already used in the past (Lee and Fassois, 1993; Fassois and Lee, 1993) and in Sec. 7 the method is concluded and some remarks for further research are outlined.

2 THE VAR(n) STRUCTURE

2.1 The Model

Let $\mathbf{Y}[t] = \begin{bmatrix} y_1[t] & y_2[t] & ... & y_s[t]\end{bmatrix}^T$ denote a s–dimensional vector time–series of zero mean random variables[1]. Under the stationarity assumption (Box et al., 2008), $\mathbf{Y}[t]$ can be described by a finite order VAR model of the following form:

$$\mathbf{Y}[t] + \mathbf{A}_1 \cdot \mathbf{Y}[t-1] + ... + \mathbf{A}_n \cdot \mathbf{Y}[t-n] = \mathbf{Z}[t] \qquad (1)$$

In the above equation n is the order of the VAR process, $\mathbf{A}_j$ designate the $[s \times s]$ AR matrices and $\mathbf{Z}[t]$ describes a vector white noise process with zero mean,

$$\boldsymbol{\mu}_\mathbf{Z} \equiv E\Big\{\mathbf{Z}[t]\Big\} = \mathbf{0} \qquad (2)$$

and covariance function,

$$\boldsymbol{\Gamma}_\mathbf{Z}[h] \equiv E\Big\{\mathbf{Z}[t+h] \cdot \mathbf{Z}^T[t]\Big\} = \begin{cases} \boldsymbol{\Sigma}_\mathbf{Z} & h = 0 \\ \mathbf{0} & h \neq 0 \end{cases} \qquad (3)$$

where $\boldsymbol{\Sigma}$ is a non–singular (and generally non–diagonal) matrix.

Taking advantage of the backshift operator q, defined such that $q^{-k} \cdot \mathbf{Y}[t] = \mathbf{Y}[t-k]$, the VAR(n) structure can be compactly written as,

$$\mathbf{A}(q) \cdot \mathbf{Y}[t] = \mathbf{Z}[t] \qquad (4)$$

where $\mathbf{A}(q)$ is the $[s \times s]$ AR polynomial matrix:

$$\mathbf{A}(q) = \mathbf{I}_s + \mathbf{A}_1 \cdot q^{-1} + ... + \mathbf{A}_n \cdot q^{-n} \qquad (5)$$

2.2 Reduction to State–space

Any VAR(n) process of Eq. 1 can be transformed to an equivalent VAR(1) structure (Lütkepohl, 2005). Define the $[n{\cdot}s \times 1]$ vectors,

$$\boldsymbol{\Xi}[t] = \begin{bmatrix} \mathbf{Y}[t-n+1] \\ \mathbf{Y}[t-n+2] \\ \vdots \\ \mathbf{Y}[t-1] \\ \mathbf{Y}[t] \end{bmatrix} \qquad \boldsymbol{N}[t] = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \\ \mathbf{Z}[t] \end{bmatrix} \qquad (6)$$

and the $[n{\cdot}s \times n{\cdot}s]$ and $[s \times n{\cdot}s]$ matrices,

$$\mathbf{F} = \begin{bmatrix} \mathbf{O}_s & \mathbf{I}_s & \dots & \mathbf{O}_s \\ \mathbf{O}_s & \mathbf{O}_s & \dots & \mathbf{O}_s \\ \dots & \dots & \ddots & \vdots \\ \mathbf{O}_s & \mathbf{O}_s & \dots & \mathbf{I}_s \\ -\mathbf{A}_n & -\mathbf{A}_{n-1} & \dots & -\mathbf{A}_1 \end{bmatrix} \qquad (7)$$

$$\mathbf{C} = \begin{bmatrix} \mathbf{O}_s & \mathbf{O}_s & \dots & \mathbf{I}_s \end{bmatrix} \qquad (8)$$

respectively, Eq. 1 can take the following form:

$$\boldsymbol{\Xi}[t] = \mathbf{F} \cdot \boldsymbol{\Xi}[t-1] + \boldsymbol{N}[t] \qquad (9)$$

$$\boldsymbol{Y}[t] = \mathbf{C} \cdot \boldsymbol{\Xi}[t] \qquad (10)$$

Equations 9–10 illustrate the state–space realization of the VAR(n) structure of Eq. 1. Naturally, the state–space model consists of a state equation (Eq. 9), in which $\mathbf{F}$ is the state transition matrix, and an observation equation (Eq. 10) that relates the original s–variate time–series $\mathbf{Y}[t]$ to the state vector, $\boldsymbol{\Xi}[t]$, by means of the output matrix $\mathbf{C}$. Obviously, the state equation can be viewed as a VAR(1) model, in which $\boldsymbol{\Xi}[t]$ is a well–defined stationary stochastic process and $\mathbf{N}[t]$ has properties similar to that of $\mathbf{Z}[t]$, as it will become clear at the following.

It must be noted that the state–space realization of Eq. 1 is not unique (Lardies, 2008). In fact they exist infinitely many pairs $\{\mathbf{F}, \mathbf{C}\}$ that can describe $\mathbf{Y}[t]$ in terms of Eqs. 9–10, since any transformation of the state vector by a non–singular $[n{\cdot}s \times n{\cdot}s]$ matrix $\mathbf{T}$ leads to new state equation, in which the transition matrix $\mathbf{T} \cdot \mathbf{F} \cdot \mathbf{T}^{-1}$ is similar to $\mathbf{F}$ and preserves its eigenvalues (Meyer, 2000). Yet, Eq. 1 has a very important property: the transition matrix $\mathbf{F}$, as defined in Eq. 7, is the block companion matrix of the polynomial matrix $\mathbf{A}(q)$ described by Eq. 5 and includes all the structural information of interest, regarding the process that generates the s–variate time–series $\mathbf{Y}[t]$.

[1] Throughout the paper, quantities in the brackets shall notate discrete–time units (or time lags, in the case of covariance functions) and hats shall notate estimators / estimates. $E\{\cdot\}$ shall notate expectation.

3 PROPERTIES OF THE STATE–SPACE REALIZATION

3.1 Noise

From Eqs. 6, 8, it holds that,

$$\mathbf{N}[t] = \mathbf{C}^T \cdot \mathbf{Z}[t] \tag{11}$$

so the mean value and the covariance matrix of $\mathbf{N}[t]$ are given by,

$$\boldsymbol{\mu}_{\mathbf{N}} \equiv E\left\{\mathbf{N}[t]\right\} = \mathbf{C}^T \cdot E\left\{\mathbf{Z}[t]\right\} = \mathbf{0} \tag{12}$$

and,

$$\begin{aligned}\boldsymbol{\Gamma}_{\mathbf{N}}[h] &\equiv E\left\{\mathbf{N}[t+h]\cdot\mathbf{N}^T[t]\right\} \\ &= E\left\{\mathbf{C}^T\cdot\mathbf{Z}[t+h]\cdot\mathbf{Z}^T[t]\cdot\mathbf{C}\right\} \\ &= \mathbf{C}^T\cdot E\left\{\mathbf{Z}[t+h]\cdot\mathbf{Z}^T[t]\right\}\cdot\mathbf{C} \\ &= \mathbf{C}^T\cdot\boldsymbol{\Gamma}_{\mathbf{Z}}[h]\cdot\mathbf{C}\end{aligned} \tag{13}$$

leading to,

$$\boldsymbol{\Gamma}_{\mathbf{N}}[h] = \begin{cases}\boldsymbol{\Sigma}_{\mathbf{N}} & h=0 \\ \mathbf{0} & h\neq 0\end{cases} \tag{14}$$

where $\boldsymbol{\Sigma}_{\mathbf{N}} = \mathbf{C}^T\cdot\boldsymbol{\Sigma}_{\mathbf{Z}}\cdot\mathbf{C}$.

3.2 State Vector

Since the state transition equation reflects the properties of an observed dynamic system, the output of which is the available s–variate time–series $\mathbf{Y}[t]$, it is desirable to obtain corresponding mathematical expressions that assess and quantify the relative information. Conventional time–series analysis usually is led to infinite, or recursive expressions for the representation / calculation of valuable quantities, such as the weighting function (referred to also as *Green function*, *process generating function*, or *transfer function*) and the covariance matrix. The analysis that follows leads to closed form representations, which reveal the spectral characteristics of the transition matrix $\mathbf{F}$.

3.2.1 The Green Function

Starting from the VAR(1) state equation,

$$\boldsymbol{\Xi}[t] = \mathbf{F}\cdot\boldsymbol{\Xi}[t-1] + \boldsymbol{N}[t] \tag{15}$$

it can be written as an infinite vector moving average,

$$\boldsymbol{\Xi}[t] = \sum_{k=0}^{\infty}\mathbf{F}^k\cdot\boldsymbol{N}[t-k] \tag{16}$$

which is a multivariate generalization of Wold's theorem (Box et al., 2008). Without loss of generality, assuming that $\mathbf{F}$ has a complete set of eigenvalues $\{\lambda_1,\lambda_2,\ldots,\lambda_{n\cdot s}\}$, it can be expressed as,

$$\mathbf{F} = \sum_{j=1}^{n\cdot s}\mathbf{G}_j\cdot\lambda_j \tag{17}$$

where $\mathbf{G}_k$ are the spectral projectors of $\mathbf{F}$ (refer to the Appendix for a brief introduction to the spectral properties of square matrices). The substitution of Eq. 17 to Eq. 16, using the fact that $\mathbf{G}_j^k = \mathbf{G}_j$ and $\mathbf{G}_i\cdot\mathbf{G}_j = 0,\ i\neq j$, yields,

$$\begin{aligned}\boldsymbol{\Xi}[t] &= \sum_{k=0}^{\infty}\left[\sum_{j=1}^{n\cdot s}\mathbf{G}_j\cdot\lambda_j\right]^k\cdot\boldsymbol{N}[t-k] \\ &= \sum_{k=0}^{\infty}\sum_{j=1}^{n\cdot s}\mathbf{G}_j\cdot\lambda_j^k\cdot\boldsymbol{N}[t-k] \\ &= \sum_{k=0}^{\infty}\mathbf{H}_{\boldsymbol{\Xi}}[k]\cdot\boldsymbol{N}[t-k]\end{aligned} \tag{18}$$

so that the coefficients of the weighting (Green) function can be expressed in a closed form as,

$$\mathbf{H}_{\boldsymbol{\Xi}}[k] \equiv \mathbf{F}^k = \sum_{j=1}^{n\cdot s}\mathbf{G}_j\cdot\lambda_j^k \tag{19}$$

in terms of the spectrum of $\mathbf{F}$. Notice that by definition (Eq. 16), $\mathbf{H}_{\boldsymbol{\Xi}}[k]$ can be viewed as the impulse response of the state difference equation, which generally has a decaying performance, characterized by a mixture of damped exponentials and cosines, as for example in vibrating systems, where the eigenvalues λ_k often appear in complex conjugate pairs. Furthermore it holds that (see Appendix):

$$\mathbf{H}_{\boldsymbol{\Xi}}[0] = \sum_{j=1}^{n\cdot s}\mathbf{G}_j = \mathbf{I} \tag{20}$$

3.2.2 The Covariance Matrix

The covariance matrix related to the Wold decomposition of the state equation is (Brockwell and Davis, 2002):

$$\boldsymbol{\Gamma}_{\boldsymbol{\Xi}}[h] = \sum_{k=0}^{\infty}\mathbf{F}^{k+h}\cdot\boldsymbol{\Sigma}_{\mathbf{N}}\cdot\left[\mathbf{F}^k\right]^T \tag{21}$$

Using Eq. 17, the following apply:

$$\begin{aligned}
\mathbf{\Gamma}_{\mathbf{\Xi}}[h] &= \\
&= \sum_{k=0}^{\infty} \left\{ \left[\sum_{j=1}^{n\cdot s} \mathbf{G}_j \cdot \lambda_j \right]^{k+h} \cdot \mathbf{\Sigma}_{\mathbf{N}} \cdot \left\{ \left[\sum_{m=1}^{n\cdot s} \mathbf{G}_m \cdot \lambda_m \right]^{k} \right\}^{T} \right\} \\
&= \sum_{k=0}^{\infty} \left\{ \sum_{j=1}^{n\cdot s} \mathbf{G}_j \cdot \lambda_j^{k+h} \cdot \mathbf{\Sigma}_{\mathbf{N}} \cdot \sum_{m=1}^{n\cdot s} \mathbf{G}_m^T \cdot \lambda_m^k \right\} \\
&= \sum_j \sum_m \mathbf{G}_j \cdot \mathbf{\Sigma}_{\mathbf{N}} \cdot \mathbf{G}_m^T \cdot \lambda_j^h \cdot \sum_{k=0}^{\infty} \lambda_j^k \cdot \lambda_m^k \\
&= \sum_j \sum_m \mathbf{G}_j \cdot \mathbf{\Sigma}_{\mathbf{N}} \cdot \mathbf{G}_m^T \cdot \lambda_j^h \cdot \frac{1}{1 - \lambda_j \cdot \lambda_m} \\
&= \sum_{j=1}^{n\cdot s} \mathbf{G}_j \cdot \mathbf{\Sigma}_{\mathbf{N}} \cdot \sum_{m=1}^{n\cdot s} \frac{\mathbf{G}_m^T \cdot}{1 - \lambda_j \cdot \lambda_m} \cdot \lambda_j^h \qquad (22)
\end{aligned}$$

Setting,

$$\mathbf{D}_j = \mathbf{G}_j \cdot \mathbf{\Sigma}_{\mathbf{N}} \cdot \sum_{m=1}^{n\cdot s} \frac{\mathbf{G}_m^T \cdot}{1 - \lambda_j \cdot \lambda_m} \qquad (23)$$

the covariance matrix can be expressed as:

$$\mathbf{\Gamma}_{\mathbf{\Xi}}[h] = \sum_{j=1}^{n\cdot s} \mathbf{D}_j \cdot \lambda_j^h \qquad (24)$$

Equation 24 has some important features. First, as become directly evident, it has the same form as the Green function. Second, it describes the covariance matrix in terms of the spectral properties of the transition matrix (plus the noise covariance), which, as already mentioned, contains all the information about the dynamics that produce the state vector and, thus, $\mathbf{Y}[t]$. This fact leads to a third crucial feature: for $h = 0$, Eq. 24 yields:

$$\mathbf{\Gamma}_{\mathbf{\Xi}}[0] = \mathbf{D}_1 + \mathbf{D}_2 + \cdots + \mathbf{D}_{n\cdot s} \qquad (25)$$

Recalling that $\mathbf{\Gamma}_{\mathbf{\Xi}}[0]$ can be treated as the multivariate equivalent of the variance (in fact its diagonal elements are the variances of each entry of $\mathbf{\Xi}[t]$), Eq. 25 can be used as a direct measure of the significance that every eigenvalue has, in the *total energy* of the vector time–series. This leads to the notion of *dispersion analysis*, originated in the work of (Lee and Fassois, 1993), in which the estimated modal characteristics of a vibrating system are qualified against some predefined thresholds. In the next section, a more practical version of Eq. 25 is presented, with respect to the estimation problem.

In the case that the correlation matrix is of interest, it can be calculated from (Box et al., 2008),

$$\mathbf{R}_{\mathbf{\Xi}}[h] = \mathbf{V}_{\mathbf{\Xi}}^{-1/2} \cdot \mathbf{\Gamma}_{\mathbf{\Xi}}[h] \cdot \mathbf{V}_{\mathbf{\Xi}}^{-1/2} \qquad (26)$$

where $\mathbf{V}_{\mathbf{\Xi}}$ is a diagonal matrix that contains the autocorrelations at zero lag:

$$\mathbf{V}_{\mathbf{\Xi}}^{-1/2} = diag\{\gamma_{11}^{-1/2}[0], \ldots, \gamma_{n\cdot s}^{-1/2}[0]\} \qquad (27)$$

3.3 Output Time–series

The previous analysis explored the advantages of the state equation and led to closed form representations for the coefficients of the Green function and the covariance matrix, which are exclusively depend on the spectrum of the transition matrix $\mathbf{F}$. Naturally, there exist strong connections between these quantities and the corresponding ones of the s–variate time–series $\mathbf{Y}[t]$. The link is just the output equation of the state–space model. The substitution of Eq. 16 to Eq. 10, under the result of Eq. 19, yields,

$$\begin{aligned}
\mathbf{Y}[t] = \mathbf{C} \cdot \mathbf{\Xi}[t] &= \mathbf{C} \cdot \sum_{k=1}^{\infty} \mathbf{F}^k \cdot \boldsymbol{N}[t-k] \\
&= \sum_{k=0}^{\infty} \mathbf{C} \cdot \mathbf{H}_{\mathbf{\Xi}}[k] \cdot \mathbf{C}^T \cdot \boldsymbol{Z}[t-k] \\
&= \sum_{k=0}^{\infty} \mathbf{H}_{\mathbf{Y}}[k] \cdot \boldsymbol{Z}[t-k] \qquad (28)
\end{aligned}$$

where:

$$\begin{aligned}
\mathbf{H}_{\mathbf{Y}}[k] &= \mathbf{C} \cdot \mathbf{H}_{\mathbf{\Xi}}[k] \cdot \mathbf{C}^T \\
&= \sum_{j=1}^{n\cdot s} \mathbf{C} \cdot \mathbf{G}_j \cdot \mathbf{C}^T \cdot \lambda_j^k \\
&= \sum_{j=1}^{n\cdot s} \mathbf{\Omega}_j \cdot \lambda_j^k \qquad (29)
\end{aligned}$$

The covariance matrix of $\mathbf{Y}[t]$ is:

$$\begin{aligned}
\mathbf{\Gamma}_{\mathbf{Y}}[h] &\equiv E\left\{ \mathbf{Y}[t+h] \cdot \mathbf{Y}^T[t] \right\} = \\
&= \sum_{k=0}^{\infty} \mathbf{H}_{\mathbf{Y}}[k+h] \cdot \mathbf{\Sigma}_{\mathbf{Z}} \cdot \mathbf{H}_{\mathbf{Y}}^T[k] \qquad (30)
\end{aligned}$$

Recalling that,

$$\mathbf{\Sigma}_{\mathbf{N}} = \mathbf{C}^T \cdot \mathbf{\Sigma}_{\mathbf{Z}} \cdot \mathbf{C} \qquad (31)$$

and

$$\mathbf{H}_{\mathbf{Y}}[k] = \mathbf{C} \cdot \mathbf{H}_{\mathbf{\Xi}}[k] \cdot \mathbf{C}^T \qquad (32)$$

the following apply:

$$\begin{aligned}
\mathbf{\Gamma}_{\mathbf{Y}}[h] &= \sum_{k=0}^{\infty} \mathbf{C} \cdot \mathbf{H}_{\mathbf{\Xi}}[k+h] \cdot \mathbf{C}^T \cdot \mathbf{\Sigma}_{\mathbf{Z}} \cdot \mathbf{C} \cdot \mathbf{H}_{\mathbf{\Xi}}^T[k] \cdot \mathbf{C}^T \\
&= \mathbf{C} \cdot \left\{ \sum_{k=0}^{\infty} \mathbf{H}_{\mathbf{\Xi}}[k+h] \cdot \mathbf{\Sigma}_{\mathbf{N}} \cdot \mathbf{H}_{\mathbf{\Xi}}^T[k] \right\} \cdot \mathbf{C}^T \\
&= \mathbf{C} \cdot \mathbf{\Gamma}_{\mathbf{\Xi}}[h] \cdot \mathbf{C}^T \qquad (33)
\end{aligned}$$

Thus, from Eq. 24,

$$\mathbf{\Gamma}_{\mathbf{Y}}[h] = \sum_{j=1}^{n\cdot s} \mathbf{Q}_j \cdot \lambda_j^h \qquad (34)$$

where,

$$\mathbf{Q}_j = \mathbf{C}\cdot\mathbf{D}_j \cdot \mathbf{C}^T \tag{35}$$

while if the correlation matrix $\mathbf{R}_\mathbf{Y}[h]$ is needed, corresponding versions of Eqs. 26– 27 apply to Eq. 34 as well.

Equations 29 and 34 show how the properties of the s–variate time–series $\mathbf{Y}[t]$ are related to the transition matrix. It is important to observe that the above analysis is strictly depended on the spectrum of $\mathbf{F}$. Indeed, when the VAR(n) structure described by Eq. 1 is available, all the information about the dynamics of the system that produces $\mathbf{Y}[t]$ can be assessed by the eigenvalue problem of $\mathbf{F}$, the state transition matrix of the state–space realization, which is identical to the block companion matrix of the polynomial matrix $\mathbf{A}(q)$. Of course, no VAR structure exists a–priori for an available data set and it rather has to be estimated. This is the topic of the next Section.

4 ESTIMATION

The estimation of VAR(n) structures pertains to the identification of the polynomial matrix order and coefficients, as well as the covariance matrix of the vector noise sequence, given observations of a s–variate times–series $\mathbf{Y}[t],\ t = 1,...,N$, that has been sampled at a period T_s. To this, the state–space realization may again be utilized, noting that, regardless the selected order n of the original VAR structure, the state equation retains the first order VAR form. In addition, Eq. 15 can be written as a linear regression,

$$\mathbf{\Xi}[t] = \mathbf{\Phi}[t]\cdot \boldsymbol{f} + \mathbf{N}[t] \tag{36}$$

with,

$$\mathbf{\Phi}[t] = -\mathbf{\Xi}^T[t-1] \otimes \mathbf{I}_{n\cdot s} \quad [n\cdot s \times n\cdot s^2] \tag{37}$$

$$\boldsymbol{f} = vec\{\mathbf{F}\} \quad [n\cdot s^2 \times 1] \tag{38}$$

where $\otimes$ denotes Kronecker's product and $vec\{\cdot\}$ the vector that is produced by stacking the columns of the relative matrix, one underneath the other. Introducing,

$$\mathbf{\Xi} = \left[\mathbf{\Xi}^T[1] \ \ldots \ \mathbf{\Xi}^T[N]\right]^T \quad [N\cdot n\cdot s \times 1] \tag{39}$$

$$\mathbf{\Phi} = \left[\mathbf{\Phi}[1] \ \ldots \ \mathbf{\Phi}[N]\right]^T \quad [N\cdot n\cdot s \times n\cdot s^2] \tag{40}$$

$$\mathbf{N} = \left[\mathbf{N}^T[1] \ \ldots \ \mathbf{N}^T[N]\right]^T \quad [N\cdot n\cdot s \times 1] \tag{41}$$

the minimization of the quadratic norm,

$$V(\boldsymbol{f}) = \frac{1}{2}\cdot\mathbf{N}^T\cdot\mathbf{\Lambda}\cdot\mathbf{N} \tag{42}$$

where $\mathbf{N} = \mathbf{\Xi} - \mathbf{\Phi}\cdot \boldsymbol{f}$ and $\mathbf{\Lambda}$ any arbitrary weighting matrix (the covariance of the residual vector $\mathbf{N}$ is presently utilized, calculated as $\mathbf{I}_N \otimes \widehat{\mathbf{\Sigma}}_\mathbf{N}^{-1}$), leads to the well-known normal equations for the least–squares estimation of $\boldsymbol{f}$,

$$\mathbf{\Phi}^T\cdot\mathbf{\Phi}\cdot\widehat{\boldsymbol{f}} = \mathbf{\Phi}^T\cdot\mathbf{\Xi} \tag{43}$$

whereas the covariance matrix associated with the estimate of Eq. 43 is:

$$\mathbf{P} = \left[\mathbf{\Phi}^T\cdot\mathbf{\Lambda}\cdot\mathbf{\Phi}\right]^{-1} \tag{44}$$

The diagonal entries of $\mathbf{P}$ are the variances of the parameter vector $\boldsymbol{f}$. Thus, assuming normality (provided that $N \gg n\cdot s^2$), the 95% confidence limits are derived from $\widehat{f}_j \pm 1.96\cdot\sigma_j$ for $j = 1,\ldots,n\cdot s^2$. Note that if the zero value is contained in this interval, the relative parameter can be regarded as zero.

Having the state equation estimated, the transition to the original VAR(n) structure is designated by the matrix $\mathbf{C}$ of the state–space realization's output equation. To this, the transformation methods that were implied in Sec. 3 are applied.

5 VALIDATION

The vector time–series fitting strategy consists of finding an appropriate estimate of the order n, as well as of exploring the properties of the innovations, $\mathbf{Z}[t]$. Both may be qualified via minimization of the Bayesian Information Criterion (*BIC*), defined as,

$$BIC = ln\ det\ |\widehat{\mathbf{\Sigma}}_\mathbf{Z}| + n\cdot s^2\frac{ln\ N}{N} \tag{45}$$

while the innovations can be further tested for whiteness, using standard hypothesis tests. See (Papakos and Fassois, 2003) for details.

Once the final model has been available, complete structural information can be assessed in terms of the estimated transition matrix. Towards this, the spectrum of $\widehat{\mathbf{F}}$ is calculated, namely the eigenvalues and the eigenvectors, while using Eq. 34, the relative importance of each structural mode, within the total energy of the system is evaluated. With respect to the discussion that took place in Sec. 3.2.2, regarding the notion of the dispersion analysis, setting $h = 0$ in Eq. 34 yields:

$$\mathbf{\Gamma}_\mathbf{Y}[0] = \mathbf{Q}_1 + \mathbf{Q}_2 + \cdots + \mathbf{Q}_{n\cdot s} \tag{46}$$

Let γ_{ij} be the $[i,j]$ element of $\mathbf{\Gamma}_\mathbf{Y}[0]$. Then,

$$\gamma_{ij} = q_{1\,ij} + q_{2\,ij} + \cdots + q_{n\cdot s\,ij} \tag{47}$$

and since the eigenvalues of the transition matrix usually come as a mixture of real and complex conjugate

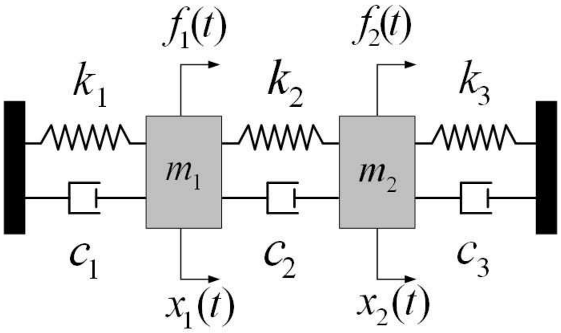


Figure 1: A structural system with two degrees of freedom: $m_1 = m_2 = 4.5\ kg,\ c_1 = 45\ Ns/m,\ c_2 = 35\ Ns/m,$ $c_3 = 15\ Ns/m,\ k_1 = k_3 = 17500\ N/m,\ k_2 = 100\ N/m$

numbers, the structural dispersions within the content of the $[i, j]$ covariance estimate are defined as:
Real mode:

$$\delta_{ij\,k} = q_{ij\,k} \tag{48}$$

Complex mode:

$$\delta_{ij\,k} = q_{ij\,k} + q^*_{ij\,k} \tag{49}$$

Thus, the relative importance of each dispersion in the $[i, j]$ covariance estimate is:

$$\Delta_{ij\,k} = \frac{\delta_{ij\,k}}{\gamma_{ij}} \times 100\% \tag{50}$$

This procedure allows the determination of the contribution of the k^{th} identified mode in every element of the covariance matrix, by building corresponding $\mathbf{\Delta}_k$, which store the relative normalized dispersions $\Delta_{ij\,k}$.

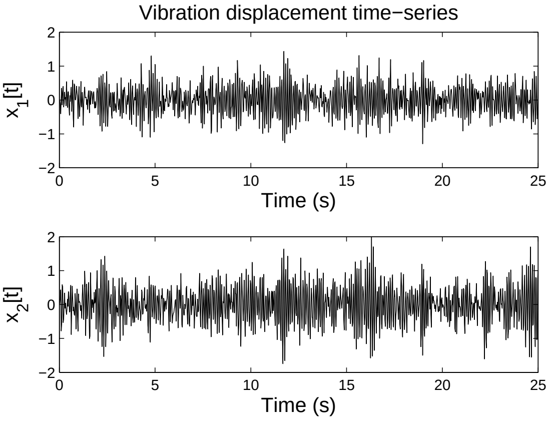


Figure 2: A realization of the noise corrupted (at 5% N/S ratio) vibration displacement time–series.

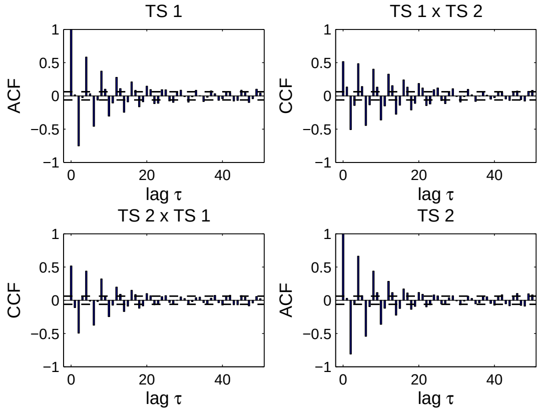


Figure 3: The correlation matrix of the series in Fig. 2 for 50 lags. TS1: $x_1[t]$, TS2: $x_2[t]$, ACF: autocorrelation, CCF: cross–correlation.

6 EXPERIMENTAL VALIDATION

The method's performance was examined through the structural identification problem of a vibrating system with two degrees of freedom, presented in Fig. 1. The system is characterized by a pair of closely spaced modes, as indicated in Tab. and the vector time–series used for the identification tasks was the vibration displacement of the masses. The statistical consistency of the method was investigated via Monte Carlo analysis that consisted of 20 data records of vibration displacement time–series (with each such record having 1000 samples, see Fig. 2 for a single realization and Fig. 3 for its covariance matrix), obtained with different white excitations and noise–corrupted at 5% noise to signal (N/S) ratio. Regarding the simulation, the continuous system was discretized using the impulse–invariant transformation, at a sampling period $T_s = 0.025\ s$.

Following the estimation procedure described in Sec. 5, a VAR(2) structure was found adequate to describe system dynamics. Table 1 illustrates the estimates of the natural frequencies and the damping ratios (in fact the corresponding mean values and the standard deviations of the Monte Carlo simulation), together with the theoretical ones, from where it is clear that the method performed satisfactory and identified the relative quantities, even in the absence of the input excitations. Table 1 further displays the percentage dispersion matrices for each mode of vibration, showing that the second mode is a heavier contributor in the total energy of the system. This result coincides with the previous assessment of the specific simulated system, reported in (Fassois and Lee, 1993).

For further validation of the results, Figs. 4–5 display the theoretical correlation matrix of the estimated model for the vector time–series of Fig. 1 and the sample correlation matrix of the innovations, for the

Table 1: Theoretical / identified natural frequencies (Hz) and damping ratios and dispersions of the identified VAR(n) structure.

			Dispersion matrices (%)	
	Theoretical	Identified	$1^{st}Mode$	$2^{nd}Mode$
w_n	9.9793	9.9903 ± 0.1446		
	9.9274	9.9307 ± 0.0636	$\begin{bmatrix} 37.68 \pm 5.06 & -32.95 \pm 8.82 \\ -37.01 \pm 10.18 & 9.58 \pm 2.73 \end{bmatrix}$	$\begin{bmatrix} 62.32 \pm 5.06 & 132.95 \pm 8.82 \\ 137.01 \pm 10.18 & 90.42 \pm 2.73 \end{bmatrix}$
ζ	0.1826	0.1848 ± 0.0174		
	0.0480	0.0477 ± 0.0073		

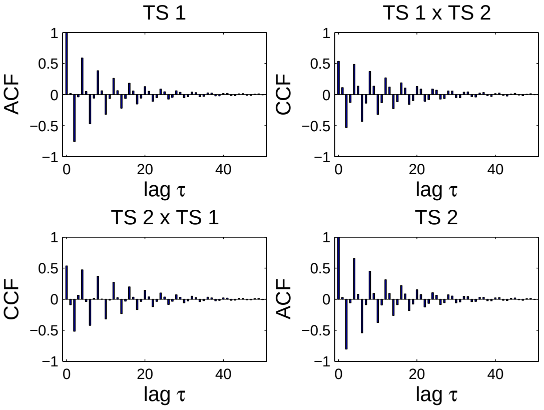


Figure 4: Theoretical correlation matrix: estimated model for the series in Fig. 2 (50 lags). Notation is the same as in Fig. 3.

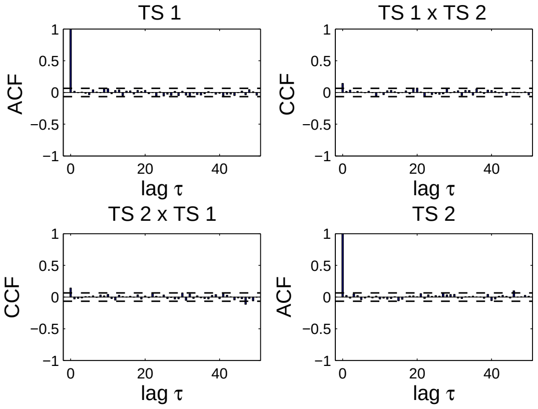


Figure 5: Sample correlation matrix: innovations of the estimated model for the series in Fig. 2 (50 lags). Notation is the same as in Fig. 3.

same model. The estimated theoretical correlation is very accurate, it follows its sample counterpart and exhibits a damped sinusoidal behavior, as a result of the identified complex conjugate eigenvalues of the transition matrix. In addition, the sample correlations of the innovations satisfy the whiteness hypothesis test, at a 95% level of significance, since they are kept within the $1.96/\sqrt{N}$ thresholds (Fig. 5, dash lines).

7 CONCLUSIONS

A novel method for the representation of vector time–series, by means of VAR(n) structures, was presented in this paper. Focusing on the estimation of structural information, the method takes advantage of the fact that every VAR(n) structure can be turn into a VAR(1) counterpart and is led to a state–space realization, whose transition matrix coincides with the block companion matrix of the VAR polynomial. Consequently, it is shown how important quantities of the original VAR(n) structure, such as the Green function and the covariance matrix, can be qualified and assessed only in terms of the spectrum of the transition matrix. This fact provides the user with the ability to accurately evaluate the significance of every structural mode in the total vector time–series energy (a technique referred to as dispersion analysis). Of the advantages of the method is the avoidance of iterative iteration schemes and the estimation of a unique structure for a given data set.

The encouraging results (reduced data acquisition, statistical consistency, accurate structural identification, no overdetermination, unique estimate) suggest the further research into this field. Extension of the method to vector time–series with structural indices governed by multiple eigenvalues, probably by means of Jordan canonical forms, as well as the investigation of VARMA models, ensues straightly. Of main interest is also the application of the method under the availability of input excitation and the expansion of its framework to non–stationary vector time–series, to closed–loop operations, as well as to fault diagnosis schemes.

REFERENCES

Box, G., Jenkins, G., and Reinsel, G. (2008). *Time Series Analysis, Forecasting and Control*. Prentice–Hall International, New Jersey, 4^{th} edition.

Brockwell, P. and Davis, R. (2002). *Introduction to Time Series and Forecasting*. Springer–Verlag, New York.

Clements, M. and Henry, D. (1998). *Forecasting Economic Time Series*. Cambridge University Press, Cambridge.

Dertimanis, V. (2006). *Fault Modeling and Identification in Mechanical Systems*. PhD thesis, School of Mechanical Engineering, GR 157 80 Zografou, Athens, Greece. In greek.

Fassois, S. and Lee, J. (1993). On the problem of stochastic experimental modal analysis based on multiple–excitation multiple–response data, Part II: The modal analysis approach. *Journal of Sound and Vibration*, 161(1):57–87.

Hannan, E. (1976). The identification and parameterization of ARMAX and state space forms. *Econometrica*, 44(4):713–723.

He, J. and Fu, Z.-F. (2001). *Modal Analysis*. Butterworth–Heinemann, Oxford.

Huang, C. (2001). Structural identification from ambient vibration using the multivariate AR model. *Journal of Sound and Vibration*, 241(3):337–359.

Koulocheris, D., Dertimanis, V., and Spentzas, C. (2008). Parametric identification of vehicle structural characteristics. *Forschung im Ingenieurwesen*, 72(1):39–51.

Lardies, J. (2008). Relationship between state–space and ARMAV approaches to modal parameter identification. *Mechanical Systems and Signal Processing*, 22(3):611–616.

Lee, J. and Fassois, S. (1993). On the problem of stochastic experimental modal analysis based on multiple–excitation multiple–response data, Part I: Dispersion analysis of continuous–time systems. *Journal of Sound and Vibration*, 161(1):57–87.

Ljung, L. (1999). *System Identification: Theory for the User*. Prentice–Hall International, New Jersey, 2 nd edition.

Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer–Verlag, Berlin.

Meyer, C. (2000). *Matrix Analysis and Applied Linear Algebra*. Society for Industrial and Applied Mathematics, Philadelphia.

Papakos, V. and Fassois, S. (2003). Multichannel identification of aircraft skeleton structures under unobservable excitations: A vector AR/ARMA framework. *Mechanical Systems and Signal Processing*, 17(6):1271–1290.

APPENDIX: MATRIX SPECTRUM

Every $[n \times n]$ matrix $\mathbf{A}$ with spectrum,

$$\sigma(\mathbf{A}) = \{\lambda_1, \lambda_2, ..., \lambda_k\}, \;\; k \leq n \tag{51}$$

has the following properties (Meyer, 2000):

- It is similar to a diagonal matrix.
- It retains a complete linearly independent set of eigenvectors.
- Every λ_j is semi–simple.

Any such matrix can be written as,

$$\mathbf{A} = \lambda_1 \cdot \mathbf{G}_1 + \lambda_2 \cdot \mathbf{G}_2 + ... + \lambda_k \cdot \mathbf{G}_k \tag{52}$$

where the $\mathbf{G}_j$'s are the, so called, spectral projectors, for which the following properties hold:

▷ $\mathbf{G}_1 + \mathbf{G}_2 + ... + \mathbf{G}_k = \mathbf{I}$

▷ $\mathbf{G}_i \cdot \mathbf{G}_j = 0, \;\; i \neq j$

▷ $\mathbf{G}_i^m = \mathbf{G}_i$

There are various ways to calculate the spectral projectors. Among them, the one that is presently utilized uses only the matrix $\mathbf{A}$ and its eigenvalues λ_j to compute $\mathbf{G}_j$:

$$\mathbf{G}_j = \frac{\prod\limits_{\substack{i=1 \\ i \neq j}}^{k} \left(\mathbf{A} - \lambda_i \cdot \mathbf{I}\right)}{\prod\limits_{\substack{i=1 \\ i \neq j}}^{k} \left(\lambda_j - \lambda_i\right)} \tag{53}$$

STRAIN FIELD INTERPOLATION OVER THE SCIARA DEL FUOCO (STROMBOLI VOLCANO) FROM GEODETIC MEASUREMENTS ACQUIRED BY THE AUTOMATIC THEODOROS SYSTEM

Giuseppe Nunnari, Alessandro Spata
Dipartimento di Ingegneria Elettrica, Elettronica e dei Sistemi, Universitá degli Studi di Catania
Viale A. Doria 6, 95125 Catania, Italy
gnunnari@dees.unict.it, alessandro_spata@yahoo.it

Giuseppe Puglisi, Alessandro Bonforte, Francesco Guglielmino
Istituto Nazionale di Geofisica e Vulcanologia, sezione di Catania, Piazza Roma 2, 95125 Catania, Italy
puglisi-g@ct.ingv.it, bonforte@ct.ingv.it, guglielmino@ct.ingv.it

Keywords: Automatic monitoring system, Data-processing, Strain interpolation, Stromboli volcano.

Abstract: In this paper we treat two important aspects concerning the automatic monitoring of the ground deformation at the Sciara del Fuoco (SdF), the Stromboli volcano (Italy) steep flank subjects to dangerous landslide events: the developments of suitable software procedures to process observations and the evaluation of both 3D motion maps and 3D strain tensor over the whole investigated area. Ground deformation measured by the monitoring system known as THEODOROS (THEOdolite and Distancemeter Robot Observatory of Stromboli) is often affected by offsets and spikes, due to both malfunctioning and periodical maintenances of the system, and other noise sources making very difficult the interpretation of the ground deformation dynamics. To this purpose a suitable software tool able to reduce these drawbacks was developed. Furthermore, both 3D motion maps and 3D strain tensor are computed in order to provide new useful information aimed to better understanding the complex dynamic of the SdF.

1 INTRODUCTION

Stromboli is an active volcano, about 2500 m high above the sea floor. Roughly only the last kilometre of this volcano emerges from the sea, forming an island whose diameter ranges from 2.4 to 5 km. It belongs to the Aeolian Islands and represents the most active volcano of this archipelago. Its conic shape is evidently characterized by a big depression that marks the northwestern flank of the edifice: the Sciara del Fuoco (SdF). On December 28th, 2002, lava flows outpoured from the northern wall of the NE crater and descended into the Sciara del Fuoco, a deep depression marking the NW flank of the volcano edifice. On December 30th, 2002, two landslides occurred on the northern part of the Sciara del Fuoco; they moved a mass in the order of tens of millions of cubic meters both above and below sea level. The landslide produced a tsunami causing significant damage to the eastern coast of the island, reaching the other Aeolian Islands and the Sicilian and southern Italian coasts. This event led to the upgrading of the ground deformation monitoring system, already existing on the island; the new requirement was the real-time detection of the deformations related to potential slope failures of the SdF. To this purpose the chosen instrument was the Leica TCA 2003 Total Station (TS) equipped with GeoMos software ((Leica, 2002) that allows remote sensor control. The acronym of this system is THEODOROS (THEOdolite and Distancemeter Robot Observatory of Stromboli) (Puglisi et al., 2004). The rest of this paper is organized in the following way. In Sec. II a brief description of the current THEODOROS configuration is given, the interested reader can found more detailed information and a general map of the island with the position of the reflectors in (Puglisi et al., 2004) and (Nunnari et al., 2008). Sec. III reports the approach adopted to preprocessing data; Sec. IV shows the methodology used to compute the strain field; Sec. V reports the case study; finally Sec. VI draws the conclusions of this study.

2 BRIEF INTRODUCTION TO THE THEODOROS SYSTEM

The THEODOROS system consists of a remote-controlled Total Station that can be programmed to measure slope distances and angles between the sensor and benchmarks appropriately installed in the SDF area at a specific sampling rate. To ensure a continuous stream of data from the instrument, it requires a constant power supply and a continuous link with the PC controlling the Total Station's activities, installed on the S. Vincenzo Observatory, where the National Department of Civil Protection (DPC) control room is located. The Stromboli volcano eruption of the 27 February 2007 destroyed the THEODOROS benchmarks inside the SDF. A new configuration was designed and new benchmarks were installed on the new fan produced by the lava flow entering the sea. This new topology consists of six reflectors installed outside the SdF, around the Total Station, for the reference system and atmospheric corrections (SENT, BORD, SEMF, SPLB2, CIST and ELIS), nine reflectors for monitoring movements of the lava fan inside the SdF (SDF18, SDF19, SDF20, SDF21, SDF22, SDF23, SDF24, SDF25 and SDF26), two reflectors to monitor the northern border of the SdF (400 and BASTI) and two further reflectors on stable sites to check the stability of the measurements both on short and very long distance measurements (CURV and CRV). Currently the reflectors SDF20 and SDF21 are not working. The sample time indicated as t_c hereafter is set to be $t_c = 10$ minutes. Each measurement for each target or reference point provides the instantaneous values of three relevant pieces of information: the slope distance (sd), the horizontal (hz) angle and the vertical angle (ve). Starting from this information, the GeoMos system is able to transform the TS measurement vectors (whose components are sd, hz, ve) into an equivalent vector whose components are expressed in terms of North (N), South (S) and Up (U) with respect to the assumed reference system. In this computation, GeoMos is able to take into account the constraints imposed by the assumption of the reference system. Despite the availability of real-time information, this is not enough to automatically evaluate the state of ground deformation. Indeed the acquired measures are affected by offsets, spikes and noise sources that strongly compromised their interpretation. These drawbacks must be necessary overcome before that suitable quantities related to the ground deformation dynamic can be efficiently computed. In particular in this paper we focus our attention on the problems of offsets and spikes removal, smoothing noisy data and strain tensor evaluation.

3 PRE-PROCESSING DATA

The algorithm we propose to remove both spikes and offsets consists of two steps. First the spikes are removed, then attention is focused on offsets. Since the single displacement components (North, East, Up) of each benchmark in the period June 2006 - December 2008 are characterized by a normal distribution, the problem to remove the spikes affecting observations, i.e. the sharp variations of the time series which are generally due to either periodical maintenance or instrument malfunctions, is well solved adopting the standard deviation of observations as reference. Indeed let $T_{SDF_x}(t)$ be a generic component of the benchmark SDF_x at time t, let $\Delta T_{SDF_x}(t)$ be the difference between two subsequent measures and denoting as σ its standard deviation, the experience gained through the daily monitoring of the SdF suggest us to consider as spikes the $\Delta T_{SDF_x}(t)$ values falling outside the range covered by one σ.

The offsets affecting observations are essentially due to the maintenances of the THEODORO system. Here it is necessary to distinguish two types of maintenances: periodical maintenance usually carried out every six months, and extra maintenance due to unexpected crash of the system. The offsets related to the periodical maintenance are simply adjusted taking into account the marked sharp variation (jump) visible when the system begins to work. This approach is also suitable for offset due to the crash of the system if the normal functioning of the system is promptly restored. Instead, if the extra maintenance is performed after a sufficiently long time the system crashed, then the offsets removal is not trivial. Indeed, in this case, in order to perform a reliable offsets correction the estimation of the trend of each ground deformation component during the period in which the system was crashed is needed. In order to adjust these kinds of offsets we use the linear trend as shown in figure 2.

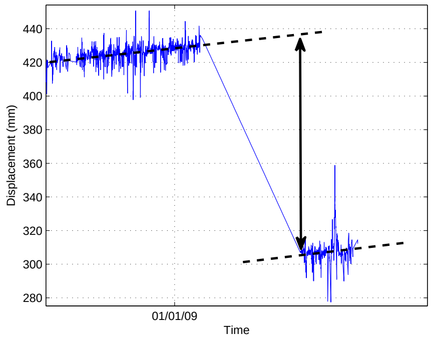


Figure 1: Offsets correction approach based on linear trend.

Although both spikes and offsets removal makes ground deformations more readable, further processing is needed to reduce noise source affecting data, in particular the thermoelastic effects on ground deformation due to the temperature. To this purpose we have smoothed noisy data with spline functions following the suggestion of the literature (Biloti et al., 2008), (Ge et al., 2003). A spline function $s(t)$ is a function defined piecewise by polynomials. This function takes values from an interval $[a,b]$ and maps them to R, the set of real numbers. The interval $[a,b]$ is divided into k disjoint subintervals $[t_i, t_{i+1}]$ with $0 \leq i \leq k-1$ such that $[a,b] = [t_0,t_1]U \dots U[t_{k-2},t_{k-1}]$. The given k points t_i are called knots. The vector $t = (t_0, .., t_{k-1})$ is called a knot vector for the spline. If the knots are equidistantly distributed in the interval $[a,b]$ the spline is uniform, otherwise it is non-uniform. On each of this subintervals a nth polynomial is defined and joined with others polynomials at the knot points in such a way that all derivatives up to the $(n-1)th$ degree are continuous. Within these constraints, the function $s(t)$ is selected which minimizes:

$$\sum (s(t_i) - x_i)^2 + p \int (s^{(\frac{n+1}{2})}(t))^2 dt \qquad (1)$$

where (t_i, x_i) are the raw data samples and $s(k)$ denotes the kth derivative of $s(t)$. The weight factor p is the smoothing parameter whose value must be opportunely chosen to obtain a good compromise between good fit and the smoothness. In figure 2 are shown, respectively, the raw data of the benchmark SdF26 (North component), the data after removing spikes and offsets and finally the spline-smoothing.

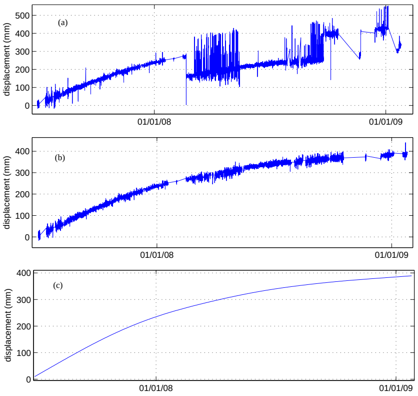


Figure 2: (a) Original SDF26 North component; (b) Spikes and offset removed; (c) Smoothing noise.

4 STRAIN INTERPOLATION

In order to compute both 3D displacements map and strain tensor in the area of SdF covered by the THEODOROS system we use the modified least-square approach introduced by (Shen et al., 1996) and also used by (Pesci and Teza, 2007) and (Teza et al., 2008). Given a point P having position $x0 = (x_{10}, x_{20}, x_{30})$ surrounded by N experimental points (EPs) whose positions and displacements are respectively $x(n) = (x_{1(n)}, x_{2(n)}, x_{3(n)})$ and $u(n) = (u_{1(n)}, u_{2(n)}, u_{3(n)})$ where $n = 1..N$, the problem of estimating both the displacements gradient tensor H and the displacement components $U_i (i = 1..3)$ of the point P, according to the infinitesimal strain theory, can be modelled in terms of the following strain interpolation equations:

$$u_{i(n)}(x) = H_{ij} \Delta x_{j(n)} + U_i \quad (i,j = 1..3) \qquad (2)$$

where $\Delta x_{j(n)} = x_{j(n)} - x_{j0}$ are the relative positions of the nth EP experimental points and the arbitrary point P and $H_{ij} = \frac{\partial u_i}{\partial x_j}$ are the elements of the displacement gradient tensor. It can be decomposed in a symmetric and an anti-symmetric part as $H = E + \Omega$, where E is the strain tensor defined as:

$$E = \frac{1}{2}(H_{ij} + H_{ji}) e_i \oplus e_j \qquad (3)$$

and Ω is the rigid body rotation tensor defined as:

$$\Omega = \frac{1}{2}(H_{ij} - H_{ji}) e_i \oplus e_j \qquad (4)$$

Here e_i is the base vector of the Cartesian reference system and $\oplus$ is the tensor product. In a compact form the undetermined system of equations (2) can be written as $Al = u$ where A is the design matrix simply derivable from equation (2), $l = [U_1\ U_2\ U_3\ \varepsilon_{11}\ \varepsilon_{12}\ \varepsilon_{13}\ \varepsilon_{22}\ \varepsilon_{23}\ \varepsilon_{33}\ \omega_1\ \omega_2\ \omega_3]$ is the vector of unknown parameters and $u = [u(1)\ u(2)\ \ u(n)]^T$ is the observation vector. Assuming a uniform strain field and re-writing the previous linear equation (4) as $Al = u + e$, where e is the residual vector which model the stochastic nature of the estimation problem, a suitable method to solve the system is the Weighted Least Squares (WLS) which gives the expression (5) as a suitable formula to estimate the unknown vector l

$$\hat{l} = (A^T W A)^{-1} A^T W u \qquad (5)$$

In (5) W is the data covariance matrix. Usually W is assumed to be diagonal (uncorrelated data), i.e. of the form

$$W = diag(\sigma^{-2}_{1(1)}, \sigma^{-2}_{2(1)}, \sigma^{-2}_{3(1)}, \dots, \sigma^{-2}_{1(n)}, \sigma^{-2}_{2(n)}, \sigma^{-2}_{3(n)}) \qquad (6)$$

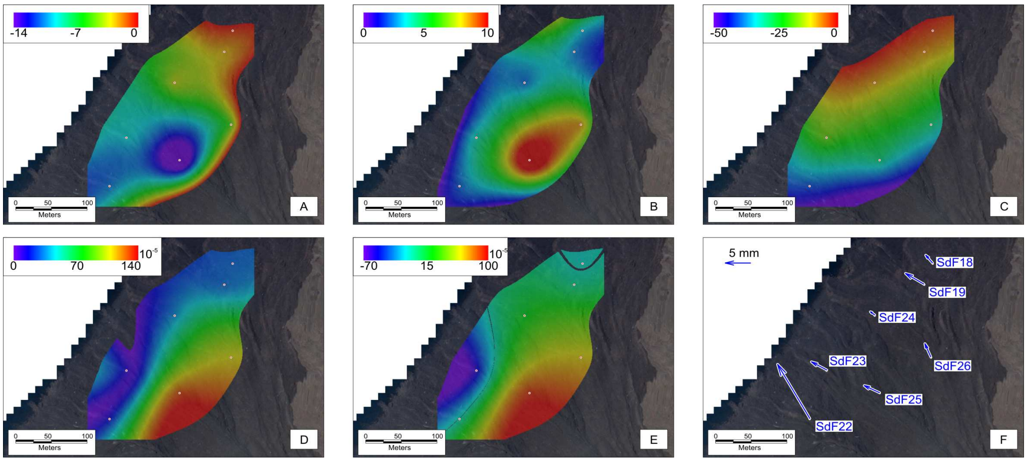


Figure 3: In the frames (a), (b), and (c) are reported the calculated East, North and Up components of displacements respectively. Frames (d) and (e) report the maximum shear strain and the volume variation. Finally, in the frame (f) the displacement vectors of the benchmarks are shown.

where the quantities $\sigma_{j(n)}$'s are the standard deviations of the measurements. According to the modified least squares (MLS) approach proposed by (Shen et al., 1996), based on the adjustment of the covariance matrix W, we use the matrix W' which is a weighted version of the matrix W of experimental data. Following the suggestion given by (Shen et al., 1996) and (Teza et al., 2008) the weight function considered here is:

$$W' = We^{-\frac{d_{(n)}}{d_0}} \tag{7}$$

where $d_{(n)}$ is the distance between the nth EP and the arbitrary point P, and d_0 is a distance-decaying constant defining the ”level of locality” of the estimation.

This method, likewise most previous methods (Frank, 1966) and (Prescott, 1976) is used to interpolate the strain among benchmarks of geodetic networks where ground deformations are measured by comparing geodetic surveys.

5 AN APPLICATION TO THE SCIARA DEL FUOCO

The benchmarks placed on the lava fan show a general NW-ward motion following the maximum slope of the SdF, with an increasing magnitude from NE to SW (Fig. 3f). This kind of deformation is in good agreement with a seawards motion of the new lava fan, driven by a mainly gravitational dynamics. However, the ground motion is not uniform above the investigated area, showing significant differences in the displacements measured on different benchmarks. In order to analyze the ground deformation pattern recorded from December 14, 2008 to January 3, 2009 above the deforming lava body, we performed a strain interpolation. Unfortunatly the corresponding linear system is undetermined since it implies more unknows ($n = 9$) that equations ($m = 7$). Therefore the solution is never unique. To this reason we have calculated a basic solution with almost m non zero components by using the QR factorization with column pivoting. Results are reported in Fig. 3, where the decrease of the horizontal motion (Fig. 3a and b) is evident from benchmark SDF25, that is located on the upper and central area of the fan, towards the coastline and towards North, reaching the minimum values at SDF18 benchmark, located close to the SdF northern rim. The vertical motion (Fig. 3c) shows a more uniform gradient, from a maximum down-lift of about 50 mm at the S-Westernmost benchmark (SDF22) to near 0 at SDF18. A deeper analysis of the overall deformation of the lava fan is allowed by the interpolation of the strain tensor. In Fig. 3d, the distribution of the maximum 3D shear strain is reported, confirming the strongest deformation on the upper area of the lava fan; this is mainly due to the stronger magnitude of horizontal displacements of the southernmost SDF22, SDF25 and SDF26 benchmarks with respect to the northern half of the fan, but also to the relative vertical motion of the two halves of the body. On the upper area, also the volumetric dilatation evidences a maximum expansion (Fig. 3e), mainly imputable to the divergent directions of the displacements affecting SDF25 and SDF26 benchmarks. In addition, a contracting area is detected on the southern coastline of the fan, due to the smaller displacements of the SDF23 and SDF24 benchmarks with respect to the upper ones, while all the northern half of the lava body shows no significant volumetric strain variation confirming the higher stability of this portion of the fan that is buttressed by the stable northern wall of the SdF.

6 CONCLUSIONS

In this paper we have first shown the pre-processing techniques adopted to reduce noise sources affecting ground deformation measures acquired at the Sciara del Fuoco by the automatic monitoring system referred to as THEODOROS. In particular, due to the gaussian distribution of acquisitions, the problem of spikes removal was simply solved taking into account their standard deviations. The offsets due to the crash of the system have been adjusted based on the evaluation of the linear trend of observations. Finally spline functions have been used to reduce the thermoelastic effects. After these pre-processing steps we have shown the based on infinitesimal strain theory method used to compute both displacements maps and strain field over the area covered by the THEODOROS system. Finally a case study related to the ground motion observed in the period December 2008 - January 2009 was carried out in order to test the proposed methodology. Preliminary results show that the distribution of the maximum 3D shear strain emphasizes the strongest deformation on the upper area of the lava fan. Furthermore the volume variation highlights a contracting area on the southern coastline of the fan. Finally all the northern half of the lava body shows no significant volumetric strain variation confirming the higher stability of this portion of the fan that is buttressed by the stable northern wall of the SdF.

REFERENCES

Biloti, R., Santos, L. T., and Martin, T. (2008). Automatic smoothing by optimal splines. *Rev. Bras. Geof*, 21(2):173–177.

Frank, F. C. (1966). Deduction of earth strains from survey data. *Bull. Seismol. Soc. Am.*, 56(1):35–42.

Ge, L., Chang, H. C., Janssen, V., and Rizos, C. (2003). The integration of gps, radar interferometry and gis for ground deformation monitoring. In *TInt. Symp. on GPS/GNSS*. UTAS.

Leica (2002). Software geomos user manual. *LEICA and GEODETICS Inc.*

Nunnari, G., Puglisi, G., and Spata, A. (2008). A warning system for stromboli volcano based on statistical analysis. *PAGEOPH*, 165(8):1619–1641.

Pesci, A. and Teza, G. (2007). Strain rate analysis over the central apennines from GPS velocities: the development of a new free software. *Bollettino di Geodesia e Scienze Affini*, 56:69–88.

Prescott, W. H. (1976). An extension of franks method for obtaining crustal shear strains from survey data. *Bull. Seismol. Soc. Am.*, 66(6):1847–1853.

Puglisi, G., Bonaccorso, A., Mattia, M., Aloisi, M., Bonforte, A., Campisi, O., Cantarero, M., Falzone, G., Puglisi, B., and Rossi, M. (2004). New integrated geodetic monitoring system at stromboli volcano (italy). *Engineering Geology*, 79(1-2):13–31.

Shen, Z. K., D., D., and Ge, B. X. (1996). Crustal deformation across and beyond the los angeles basin from geodetic measurements. *Journal of Geophysical Research*, 101(27):957980.

Teza, G., Pesci, A., and Galgaro, A. (2008). Gridstrain and Gridstrain3 : Software packages for strain field computation in 2D and 3D environments. *Computers and Geosciences*, 34(9):1142–1153.

ON MODIFICATION OF THE GENERALISED CONDITIONING TECHNIQUE ANTI-WINDUP COMPENSATOR

Dariusz Horla
Poznan Univeristy of Technology, Institute of Control and Information Engineering, ul. Piotrowo 3a, 60-965 Poznan, Poland
dariusz.horla@put.poznan.pl

Keywords: Back-calculation, Generalised conditioning technique, Anti-windup compensator, Constraints.

Abstract: New anti-windup scheme is presented in application to pole-placement control, with a complete analysis of its behaviour for a class of second-order minimumphase stable plants of oscillatory and aperiodic characteristics with different dead-times. The classical generalised conditioning technique anti-windup compensator performance is compared with its three proposed modifications, arising in a new GCT compensation scheme. A critical discussion of the necessity of compensation is also given.

1 INTRODUCTION

Constraints are ubiquitous in real-world environment. As the result of their presence or the presence of some nonlinearities in the control loops, arises the difference in between computed and applied (i.e. constrained) control signal. In such a case, the performance of the closed-loop system degrades in comparison with the performance of the linear system, when constraints are not active. Such a degradation is defined as windup phenomenon (Rundqwist, 1991; Walgama and Sternby, 1990; Walgama and Sternby, 1993).

This can be also viewed from the point of discrepancy in between internal controller states and its output. When there is no correspondence in between controller's output and its internal controller states, the controller does not have any information what the current value of the constrained control signal is, and windup phenomenon arises.

The windup phenomenon has been, at first, defined for controllers comprising integral terms, as the so-called integrator windup (Rundqwist, 1991). For such controllers, control constraints may cause excessive integration of the error signal, giving rise to longer settling of the output signal and overshoots. There are two strands in compensating windup phenomenon (in AWC, anti-windup compensation) – taking constraints into account during the design procedure of the controller or assuming the system is linear, designing the controller for the linear case, and, subsequently, imposing constraints and applying AWCs (Horla, 2007; Horla and Krolikowski, 2003a; Horla and Krolikowski, 2003b).

The simplest anti-windup compensators have been based on the idea of integrator clamping, i.e. they referred to the controllers comprising integral terms only (Visioli, 2003). The proposed AWCs avoided integration of the error signal whenever some conditions were met, e.g., the control signal saturated, or error exceeded some predefined threshold, etc.

Such an approach was simple enough to be easily implemented, but as it has been already said, applicable to some controllers only.

The advanced anti-windup compensators have been designed for the case of general controller, which input-output equation is written in the RST form. Among the proposed AWCs one can find in the literature deadbeat, generalised, conditioning technique, modified conditioning technique and generalised conditioning technique anti-windup compensators (Horla and Krolikowski, 2003a; Horla and Krolikowski, 2003b). The three latter AWCs are based on the idea of back-calculation, i.e. modification of the signal that the output signal of the plant is to track, with respect to current saturation level.

The paper focuses on the generalised conditioning technique AWC (GCT-AWC), being a compromise solution in between the simplicity of the advanced AWC and compensation capabilities of the conditioning algorithm, what will be explained later.

The main idea of the paper is to present a modification of the GCT-AWC that can arise from the idea of integrator clamping methods, and to show that it can result in better control performance than performance of the system with original GCT-AWC. The presented results refer to the research carried for a set of stable minimumphase second-order discrete-time plants and different constraint levels.

There are no remarks in the literature how to improve the performance of the GCT-AWC, apart from (Horla and Krolikowski, 2003a). The proposed method limits the number of modifications, with the same excess. By introducing the proposed modifications one can improve the performance of the most appealing AWC technique.

2 PLANT MODEL

Let the discrete-time CARMA model be given

$$A(q^{-1})y_t = B(q^{-1})u_{t-d}\,, \tag{1}$$

where y_t is the plant output, u_t is the constrained control input, $d \geq 1$ is a dead-time and:

$$\begin{aligned} A(q^{-1}) &= 1 + a_1 q^{-1} + a_2 q^{-2}\,, &(2)\\ B(q^{-1}) &= b_0 + b_1 q^{-1} &(3) \end{aligned}$$

are relatively prime. The control signal $u_t = \mathrm{sat}(v_t;\alpha)$ is the computed control signal after saturation by symmetrical cut-off function at level ±α.

3 CONTROLLER

The plant is controlled by the pole-placement controller that ensures tracking of a given reference signal r_t by the plant output y_t with given dynamics,

$$\begin{aligned} v_t &= k_R r_t - k_P y_t + k_I \frac{q^{-1}}{1-q^{-1}}(r_t - y_t) - \\ &\quad -k_D \frac{1-q^{-1}}{1-\gamma q^{-1}} y_t\,, \end{aligned} \tag{4}$$

where $k_R = rk_P, r > 0$. The above controller equation can be obtained by discretisation of a continuous-time PID controller (Rundqwist, 1991), and it can be rewritten into the RST structure (Horla and Krolikowski, 2003b)

$$R(q^{-1})v_t = -S(q^{-1})y_t + T(q^{-1})r_t\,. \tag{5}$$

Coefficients of polynomials $R(q^{-1})$, $S(q^{-1})$, $T(q^{-1})$ can be determined by solving the following Diophantine equation

$$\begin{aligned} A(q^{-1})R(q^{-1}) + q^{-d}B(q^{-1})S(q^{-1}) &= \\ = A_M(q^{-1})A_o(q^{-1})\,, \end{aligned} \tag{6}$$

where polynomials $A_o(q^{-1})$ and $A_M(q^{-1})$ are stable, and given polynomial $A_M(q^{-1})$ is of second order in the chapter.

Controller polynomials $R(q^{-1})$, $S(q^{-1})$, $T(q^{-1})$ are of order $d+nB$, nA, nA_o, respectively, and have forms as follows:

$$\begin{aligned} R(q^{-1}) &= (1-q^{-1})R'(q^{-1})\,, \\ S(q^{-1}) &= s_0 + s_1 q^{-1} + s_2 q^{-2}\,, \\ T(q^{-1}) &= k_R A_o(q^{-1})\,. \end{aligned} \tag{7}$$

From the controller equation (5), given structure $R(q^{-1})$, $S(q^{-1})$, $T(q^{-1})$ (7) and (4) it follows that:

$$\begin{aligned} s_0 &= k_P + k_D\,, \\ s_1 &= k_I - 2k_D - k_P(1+\gamma)\,, \\ s_2 &= k_D - \gamma(k_I - k_P)\,, \end{aligned} \tag{8}$$

$$\begin{aligned} A_o(q^{-1}) &= \left(1-\gamma q^{-1}\right)\left(1-q^{-1}\left(1-\frac{k_I}{k_R}\right)\right) = \\ &= 1 + a_{o1}q^{-1} + a_{o2}q^{-2}\,, \end{aligned} \tag{9}$$

where $\gamma = -\frac{b_1}{b_0}, k_R = rk_P,\ a_{o1} = \frac{k_I}{k_R} - (1+\gamma), a_{o2} = \gamma\left(1-\frac{k_I}{k_R}\right)$. As the polynomial $A_o(q^{-1})$ has to be stable, $0 < \frac{k_I}{k_R} < 2$ must hold what can be ensured by a proper choice of r.

The controller algorithm is assumed to be altered by anti-windup compensator presented in the next Section, in order to assure better control performance of the closed-loop system subject to constraints. It is to be borne in mind that the compensation is based on back-calculation, i.e., it does not require the controller to have integral terms in general.

4 GENERALISED CONDITIONING TECHNIQUE AWC

In GCT, the filtered set-point signal is conditioned, instead of the set-point r_t, and given as

$$r_{f,t} = \frac{Q(q^{-1})T_1(q^{-1})}{L(q^{-1})}r_t\,, \tag{10}$$

with

$$\begin{aligned} T(q^{-1}) &= T_2(q^{-1})T_1(q^{-1})\,, \\ Q(q^{-1}) &= q_0 + q_1 q^{-1} + \cdots + q_{nQ}q^{-nQ}\,, \\ L(q^{-1}) &= 1 + l_1 q^{-1} + \cdots + l_{nL}q^{-nL} \end{aligned} \tag{11}$$

and $T_2(0) = t_{2,0}$.

Similarly to the conditioning method (see (Horla and Krolikowski, 2003a)), the modified filtered reference signal is given by

$$r^r_{f,t} = r_{f,t} + \frac{q_0(u_t - v_t)}{t_{2,0}}\,, \tag{12}$$

and the control signal is defined as

$$\begin{aligned} v_t &= (1-Q'(q^{-1})R(q^{-1}))u_t + \frac{t_{2,0}}{q_0} r_{f,t} + \\ &+ \frac{1}{q_0}((T_2(q^{-1})L(q^{-1}) - t_{2,0})r^r_{f,t} - \\ &- Q'(q^{-1})S(q^{-1})y_t\,, \end{aligned} \tag{13}$$

where $Q'(q^{-1}) = \frac{Q(q^{-1})}{q_0}$.

The GCT method enables additional tuning of the performance by reference signal filter design. Because its parameters should correspond to model parameters, saturation level and set-point values, a special choice of parameters of the filter (10) for minimumphase second-order model is proposed (Horla and Krolikowski, 2003a). Let ρ_1 and ρ_2 denote poles of stable $A(z^{-1})$, then

$$\rho = \max(|\rho_1|, |\rho_2|)\,, \tag{14}$$

$$Q(q^{-1}) = 1 + \Big((1-\rho)^{\xi} - 1\Big)q^{-1}\,, \tag{15}$$

$$L(q^{-1}) = 1 - (1-\rho)^{\xi} q^{-1}\,, \tag{16}$$

where $0 < \xi \leq 1$ is the damping factor obtained from classical root locus theory for the second-order systems. The suggested filter (14–16) takes into consideration model parameters and set-point values only, forcing the initial values of the filtered reference signal for slow models and reducing the amplitude and rate of transients for oscillatory ones.

The inherent property of the conditioning technique is the so-called short sightedness phenomenon, resulting in consecutive resaturations of the control signal because of excessive modification of the reference signal. In order to improve the performance of the compensation three modifications will be considered as in the next Section.

5 MODIFIED GENERALISED CONDITIONING AWCS

In order to combine classical conditional integration methods that work for controllers comprising integrators with back-calculation AWC presented in the previous Section, the following three back-calculation modifications have been proposed – the modification of the filtered reference input is applied when:

M1 $|e_t| > e_1\,,$

M2 $u_{t-1} \neq v_{t-1}\,,$

M3 $u_{t-1} \neq v_{t-1}$ and $e_t u_{t-1} > 0\,,$

where e_1 is a threshold value for reference modification clamping.

By applying the modifications to the GCT-AWC one assures that modification of the filtered reference signal is performed only when necessary.

6 SIMULATED PLANTS

The simulations have been performed for a set of stable, second-order, minimumphase plants with $B(q^{-1}) = 1 + 0.5q^{-1}$ and:

- P1 type

  $$A(q^{-1}) = (1 - q^{-1}(\sigma + \omega i))(1 - q^{-1}(\sigma - \omega i))\,,$$

  where:
  $$\begin{aligned} -1 < &\ \sigma\ < 1\,, \\ -1 < &\ \omega\ < 1\,, \\ |\sigma \pm \omega i| &< 1\,, \end{aligned}$$

  what corresponds to oscillatory behaviour of the plant,

- P2 type

  $$A(q^{-1}) = (1 - q^{-1}z_1)(1 - q^{-1}z_2)\,,$$

  where:
  $$\begin{aligned} 0 < &\ z_1\ < 1\,, \\ 0 < &\ z_2\ < 1\,, \end{aligned}$$

  what corresponds to aperiodic behaviour of the plant.

The simulations have been run for square wave reference signal of period 40 samples and symmetrical amplitude ±3 with $\frac{k_I}{k_R} = 0.5, A_M(q^{-1}) = 1 - 0.5q^{-1} + 0.06q^{-2}$ and $e_1 = 3$.

In order to evaluate the quality of regulation process, the performance index is introduced

$$J = \frac{1}{N}\sum_{t=0}^{N}(r_t - y_t)^2\,, \tag{17}$$

where $N = 150$ denotes the simulation horizon.

The simulations have been performed for the same constraint hardness for each of the plants, denoted by relative constraint level α_r (i.e., the multiplicity of the minimum constraint level $\alpha_{\min} = 3\frac{|A(1)|}{|B(1)|}$ allowing asymptotic tracking). The absolute value of the constraint is $\alpha = \alpha_r\alpha_{\min}$ and changes as the plants change.

7 PERFORMANCE SURFACES

The results of the simulation tests are shown as performance surfaces. Each of the axes has been divided into 101 values, thus all simulation results refer to a grid of 101 × 101 different plants. The idea of such surfaces is as follows – let J_0 denote the value of the performance index of the control system with some plant and given α_r and no AWC. Let J_1 denote the value of the performance index of the same control system with the same plant but with classical GCT-AWC. Let J_2 denote the value of the performance index of, again, the same control system with the same plant but with modified GCT-AWC (M1, M2 or M3).

For each of the plants and constraints level the following face is plotted:

- 🟪 (magenta) $J_0 = J_1 = J_2$,
- 🟥 (red) modification is of the worst performance, $J_0 < J_1 < J_2$ or $J_1 < J_0 < J_2$, the intensity of the red level is proportional to $J_2 - J_0$ or $J_2 - J_1$,
- □ (white) modification improves the performance of the GCT-AWC, $J_0 < J_2 < J_1$,
- ■ (black) it is not worth to modify GCT, $J_1 < J_2 < J_0$, the intensity of the black level is proportional to $J_2 - J_1$,
- 🟦 (blue) modification improves the performance where GCT fails to, $J_2 < J_0 < J_1$, the intensity of the blue level is proportional to $J_0 - J_2$,
- 🟩 (green) modification is of the best performance, the intensity of the green level is proportional to $J_0 - J_2$ or $J_1 - J_2$.

8 SHOULD ONE MODIFY GCT?

The performance surfaces have been obtained for P1 and P2 type plants with different dead-times and presented in Figs 1 and 2, where consecutive rows for different dead-times refer to M1, M2 and M3.

In all the cases of P1 and $d = 1$ it is visible that all modifications can improve the performance of the GCT for slow plants, i.e., with small natural frequency, whereas in the case of M1 and M3 there is an improvement visible for such plants near stability border. In the case of M1 and M3, one can see region of the best improvement (green). By comparing the given surfaces one can say that M3 is of the best AWC performance, because of the green regions and brighter red regions than in other cases, what refers to less performance degradation.

It is not advisable to modify the GCT algorithm when the region is red, it is advisable to improve where it is white and definitely advisable when green.

In the case of $d = 3$ one can see that red regions have almost disappeared and the improvement is best in the case of M3.

For P2 type plants a performance improvement can be observed for slow plants (green) with M1 and M3. Because of the size of white and green regions one can say that the best performance is assured by M1, mainly because of the $\alpha_r > 1$, that is visibility of green regions for greater α_rs. The vast areas of red color suggest that it is inadvisable to modify the original GCT when plant is moderately slow (expressed by absolute values of its poles).

In the case of $d = 3$ because of the area of white region and brightness of the red region, it can be said that M1 is the best choice, then M2 and M3.

9 SUMMARY

It has been shown that it can be advantageous to modify the algorithm of well-known GCT-AWC in order to obtain high control performance. Such a modification can be implemented with the use of lookup table, where the information is stored what GCT algorithm should be used when plant parameters vary in time, e.g. due to aging or set-point change. A similar approach has been presented for continuous system, PID controllers and integrator clamping (Visioli, 2003).

REFERENCES

Horla, D. (2007). Simple anti-integrator windup compensators – performance analysis. *Studies in Control and Computer Science*, 32:85–102.

Horla, D. and Krolikowski, A. (2003a). Anti-windup circuits in adaptive pole-placement control. In *Proceedings of the 7th European Control Conference*.

Horla, D. and Krolikowski, A. (2003b). Anti-windup compensators for adaptive pid controllers. In *Proceedings of the 9th IEEE International Conference on Methods and Models in Automation and Robotics*, pages 575–580.

Rundqwist, L. (1991). *Anti-reset Windup for PID Controllers*. PhD thesis, Lund University of Technology.

Visioli, A. (2003). Modified anti-windup scheme for pid controllers. *IEE Proceedings-D*, 150(1):49–54.

Walgama, K. and Sternby, J. (1990). Inherent observer property in a class of anti-windup compensators. *International Journal of Control*, 52(3):705–724.

Walgama, K. and Sternby, J. (1993). On the convergence properties of adaptive pole-placement controllers with anti-windup compensators. *IEEE Transactions on Automatic Control*, 38(1):128–132.

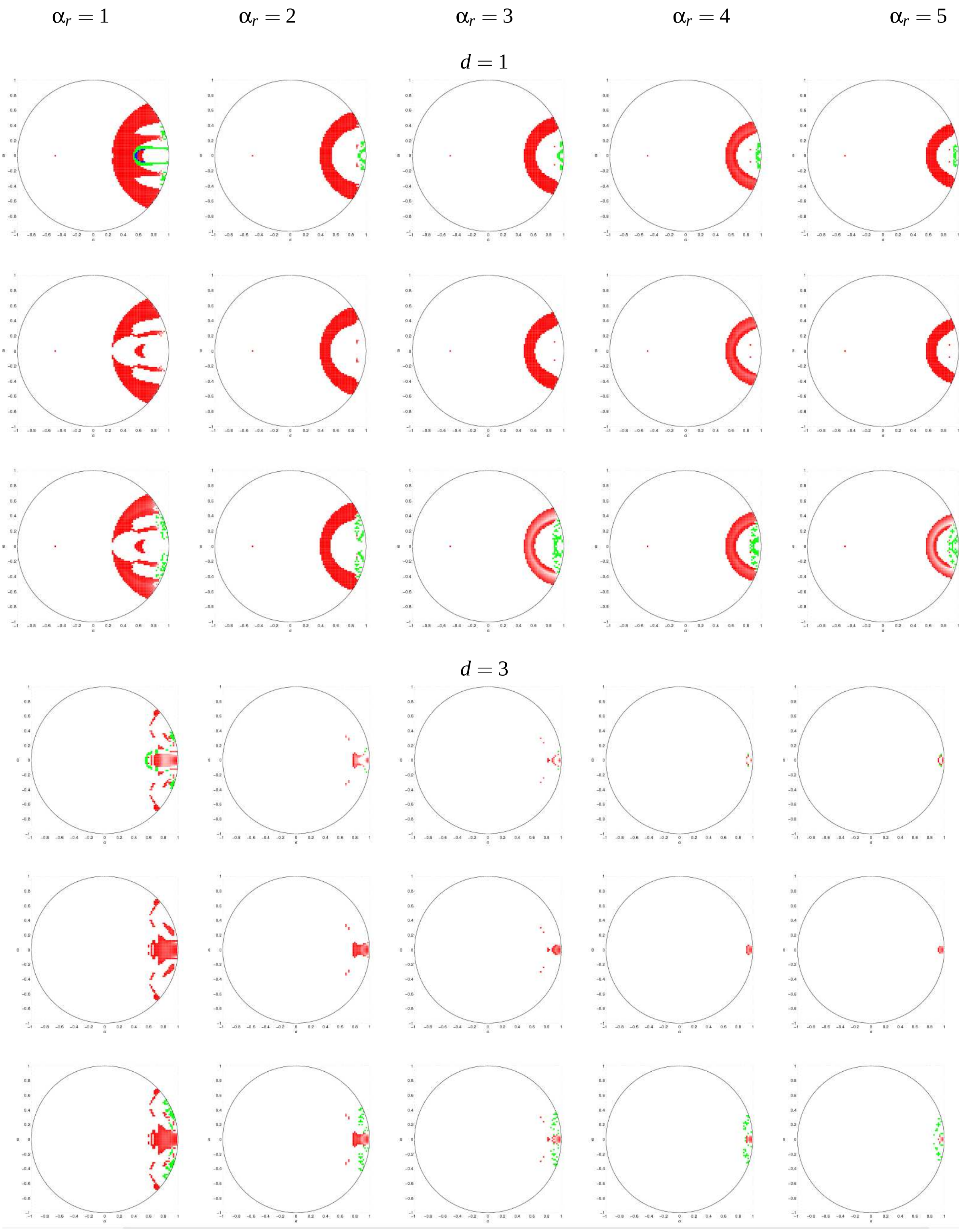


Figure 1: Performance surfaces for P1.

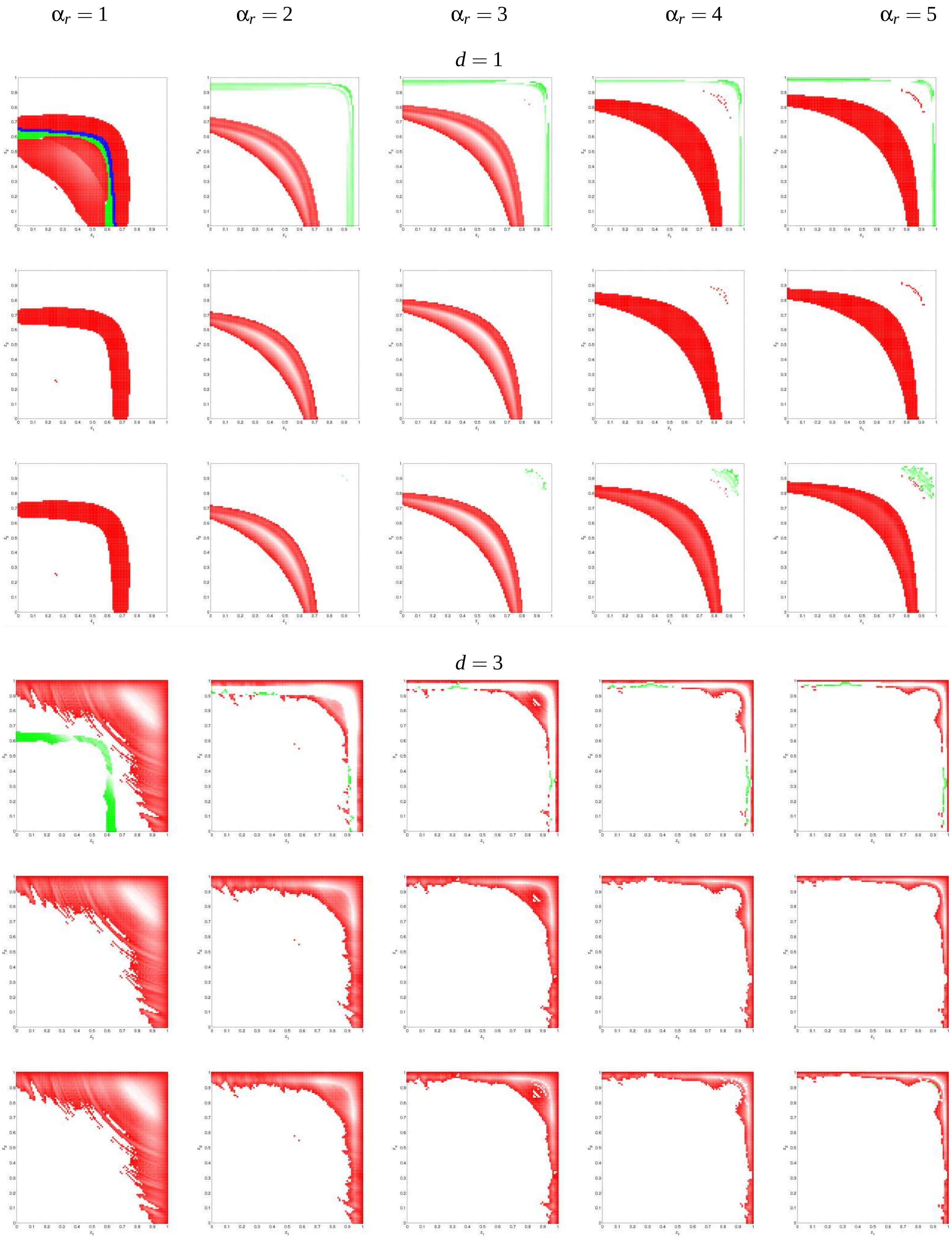


Figure 2: Performance surfaces for P2.

CO-EVOLUTION PRESERVING MODEL REDUCTION FOR UNCERTAIN CYBER-PHYSICAL SYSTEMS
Towards a Framework for Nanoscience

Manuela L. Bujorianu and Marius C. Bujorianu
School of Mathematics, University of Manchester, U.K.
Manuela.Bujorianu@manchester.ac.uk

Keywords: Cyber-physical systems, Adaptive bisimulation, Co-evolution, Stochastic model checking, Qualitative model reduction, Nanoscience.

Abstract: The problem of abstracting computational relevant properties from sophisticated mathematical models of physical environments has become crucial for cyber-physical systems. We approach this problem using Hilbertean formal methods, a semantic framework that offers intermediate levels of abstractions between the physical world described in terms of differential equations and the formal methods associated with theories of computation. Although, Hilbertean formal methods consider both deterministic and stochastic physical environments, in this paper, we focus on the stochastic case. The abstraction method can be used for verification, but also to improve the controller design and to investigate complex interactions between computation and physics. We define also a computational equivalence relation called adaptive model reduction, because it considers the co-evolution between a computation device environment and its physical environment during abstraction.

1 INTRODUCTION

The interaction between physics and computation can be very subtle. The research experience from areas like nanoscience (Hornyak e.a. 2008) and quantum computing (Accardi e.a. 2006), or from smart dust, shows that common principles can be distilled from these different worlds. At a larger scale, the general system theory provides a systematic repertoire of common properties of the physical and digital dynamical systems. This experiences give hope for a sound semantic framework for *cyber-physical systems* (CPS). The manifestos on CPS - see, for example(Tabuada 2006) - emphasize the need for a fundamentally new theoretical foundation. This foundation should be interdisciplinary and at the right level of abstraction: it should offer analytical tools to investigate physical models, and, at the same time, to be abstract enough to give semantics for models of computation.

In this paper, we consider *Hilbertian Formal Methods* (HFM) (Bujorianu, Bujorianu 2007a, 2007b) as a semantic framework for CPS modeling. HFM represent a logical framework that uses functional and stochastic analysis to construct logics for reasoning about qualitative properties of physical phenomena. These logics can be easily integrated with specification logics for automata. In this work, we focus more on the method part of HFM, and less on the formal aspects. In the HFM framework, we use hybrid systems to design an abstraction method that simplifies the physical models whilst the computational properties are simulated. Intuitively, the computational discrete steps are preserved, while the mathematical models of the continuous phenomena in the environment are drastically simplified.

The qualitative model reductions method we propose is a fundamental step towards *stochastic model checking* (SMC) (Bujorianu, Bujorianu 2006) for uncertain CPS. Stochastic model checking coincides with *probabilistic model checking* (Bujorianu, Katoen 2008) for Markov chains. In the case of continuous or hybrid stochastic dynamical systems, the SMC is a specialization of the *stochastic reachability analysis* (Bujorianu 2004) by means of computer science inspired *abstraction* (Bujorianu, Lygeros, Bujorianu 2005a) or *bisimulation methods* (Bujorianu, Lygeros, Bujorianu 2005b) , (Bujorianu, Bujorianu 2008b).

In the context of uncertain cyber-physical systems, we introduce a new concept of behavior equivalence called *adaptive bisimulation*. In the theory of concurrent discrete processes, bisimulation is a method for reducing the state space, while the transitions are preserved. Using category theory the concept of bisimulation was defined for continuous and hybrid dynamical systems (Haghverdi, Tabuada, Pappas 2005). Based on the same categorical machinery, in (Bujorianu, Lygeros, Bujorianu 2005b), bisimulation has been defined for stochastic hybrid systems. However, in the context of uncertain CPS, the classical concept of bisimulation seems to be too strong (i.e., systems that are considered equivalent by a designer or by an observer, fail to be bisimilar). More appropriate concepts of behavioral equivalence, like approximate bisimulation and behavioral bisimulation have been proposed in (Bujorianu, Bujorianu, Blom 2008) and (Bujorianu, Lygeros, Bujorianu 2005a). Under *approximate bisimulation*, the trajectories of two randomized hybrid systems differ with a small distance, the measurement being done according with a suitable metric. For the *behavioral bisimulation,* two equivalent systems have the same probabilities of reaching some specific state sets. Although these bisimulation concepts are better in describing properties of systems that operate in physical environments, they do not imply the preservation of the interaction between computation and physics. The key point in defining such a bisimulation consists in modeling this interaction. In this paper, we model this interaction using an abstract measure called *energy*, which is a basic concept of HFM. The energy characterizes globally the cyber physical process, but also it can discriminate continuous (physical) evolutions, discrete (computational) transitions and control (the process killing, in order to start another one). This last aspect makes the difference between a CPS and a classical automaton: a computation device has the capability to influence its physical environment (and achieving *co-evolution* in this way). Naturally, the CPS bisimulation should be related to energy preservation. An intuitive illustration of adaptive bisimulation is given by the following scenario. During its evolution, a CPS may produce a change of its environment. Suppose that for the new dynamical system modeling the environment is classically bisimilar with the former one. Then, for an adaptive bisimilar CPS the computational component will exhibit a equivalent behavior.

The paper road map can be described as follows. The following section contains the mathematical setting. In Section 3 we formulate the stochastic model checking problem and we prove two results that make the problem solvable. In Section 4 we investigate the qualitative model reductions and bisimulations. The final section contains some short conclusions.

2 THE MATHEMATICAL FRAMEWORK

2.1 Uncertain Cyber-physical Systems

The theory of hybrid systems is a well-established modeling paradigm for embedded systems. Similarly, the theory of concurrent embedded hybrid systems (Bujorianu, Lygeros, Bujorianu 2005a) constitutes a suitable modeling framework for CPS. In the following an uncertain cyber-physical system is modeled as a randomized embedded hybrid system.

There are two major ways to randomize a continuous or hybrid dynamical system: In one approach, the concept of noise is used to model small random perturbations. The randomized system has trajectories that closely resemble those of the deterministic initial system. The noise based randomization is carried out using stochastic differential equations. When the influence of the random perturbation changes dramatically the system evolution, the randomization is carried out using stochastic kernels that replace the concept of reset maps from deterministic hybrid system models.

A Ucps $U = (Q, \mathcal{X}, F, R, \lambda)$ consists of

- a finite set of discrete variables Q;
- a map $\mathcal{X} : Q \to \mathbb{R}^{d(\cdot)}$ that sends each $q \in Q$ into a mode (an open subset) X^q of $\mathbb{R}^{d(q)}$, where $d(q)$ is the Euclidean dimension of the corresponding mode;
- a map $F : Q \to 2^{\mathcal{F}_{SDE}}$ which specifies the continuous evolution of the automaton in terms of stochastic differential equations (SDE) over the continuous state x^q for each mode;
- a family of stochastic kernels $R = (R^q)_{q \in Q}$,

$$R^q : \overline{X}^q \times (\cup \mathcal{B}(X^j) | j \in Q \backslash \{q\}) \to [0,1];$$

- a transition rate function

$$\lambda : (\cup \overline{X}^j | j \in Q) \to \mathbb{R}^+, \tag{1}$$

which gives the distributions of the jump times.

The executions of a Ucps can be described as follows: start with an initial point $x_0 \in X^q$, follow a solution of the SDE associated to X^q, jump when this trajectory hits the boundary or according with the transition rate λ (the jump time is the minimum of the boundary hitting time and the time, which is exponentially distributed with the transition rate λ). Under standard assumptions, for each initial condition $x \in j \in Q \cup X^j$, the possible trajectories starting from x, form a stochastic process. Moreover, for all initial conditions x, the executions of a Ucps form the semantics, which can be thought of as a Markov process in a general setting. Let us consider $M = (\Omega, \mathcal{F}, \mathcal{F}_t, x_t, P_x)$ be the semantics of U. Under mild assumptions on the parameters of U, M can be viewed as a family of Markov processes with the state space $(X,\mathcal{B})$, where X is the union of modes and $\mathcal{B}$ is its Borel σ-algebra. Let $\mathcal{B}^b(X)$ be the lattice of bounded positive measurable functions on X. The meaning of the elements of M can be found in any source treating continuous-parameter Markov processes (see, for example, (Davis 1993)). Suppose we have given a σ-finite measure μ on $(X,\mathcal{B})$.

In the following we give some operator characterizations of stochastic processes, which are employed in this paper to define a qualitative model reduction for Ucps.

2.2 Hilbertean Formal Methods

The HFM abstract away the analytical properties of deterministic and stochastic differential operators using the so called kernel operator (defined in the following). Using methods of functional analysis HFM elegantly generalize both deterministic and stochastic systems. In this work we focus on the stochastic case. Let us describe briefly the mathematical apparatus that is usually employed to study continuous time continuous space Markov processes.

The *transition probability function* is $p_t(x,A) = P_x(x_t \in A)$, $A \in \mathcal{B}$. This is the probability that, if $x_0 = x$, x_t will lie in the set A.

The *operator semigroup* $\mathcal{P}$ is defined by

$$P_t f(x) = \int f(y)p_t(x,dy) = E_x f(x_t), \forall x \in X,$$

where E_x is the expectation w.r.t. P_x.

The *operator resolvent* $\mathcal{V} = (V_\alpha)_{\alpha\geq 0}$ associated with $\mathcal{P}$ is

$$V_\alpha f(x) = \int_0^\infty e^{-\alpha t} P_t f(x) dt,$$

$x \in X$. Let denote by V the initial operator V_0 of $\mathcal{V}$, which is known as the *kernel operator* of the Markov process M. The operator resolvent $(V_\alpha)_{\alpha\geq 0}$ is the Laplace transform of the semigroup.

The *strong generator* $\mathcal{L}$ is the derivative of P_t at $t = 0$. Let $D(\mathcal{L}) \subset \mathcal{B}_b(X)$ be the set of functions f for which the following limit exists (denoted by $\mathcal{L} f$):

$$\lim_{t \searrow 0} \frac{1}{t}(P_t f - f).$$

In the HFM, there is developed a semantic framework for concurrent embedded systems constructed using energy forms. We specialize this theory for function spaces, reaching in this way the theory of Dirichlet forms (Ma, Rockner 1990).

A quadratic form $\mathcal{E}$ can be associated to the generator of a Markov process in a natural way.

Let $L^2(X,\mu)$ be the space of square integrable μ-measurable extended real valued functions on X, w.r.t. the natural inner product $< f,g >_\mu = \int f(x)g(x)d\mu(x)$.

The quadratic form $\mathcal{E}$:

$$\mathcal{E}(f,g) = - < \mathcal{L} f,g >_\mu, f \in D(\mathcal{L}), g \in L^2(X,\mu) \quad (2)$$

defines a closed form. This leads to another way of parameterizing Markov processes. Instead of writing down a generator one starts with a quadratic form. As in the case of a generator it is typically not easy to fully characterize the domain of the quadratic form. For this reason one starts by defining a quadratic form on a smaller space and showing that it can be extended to a closed form in subset of $L^2(\mu)$. When the Markov process can be initialized to be stationary, the measure μ is typically this stationary distribution (see (Davis 1993) p.111). More generally, μ does not have to be a finite measure.

A *coercive closed form* is a quadratic form $(\mathcal{E},D(\mathcal{E}))$ with $D(\mathcal{E})$ dense in $L^2(X,\mu)$, which satisfies the: (i) closeness axiom, i.e. its symmetric part is positive definite and closed in $L^2(X,\mu)$, (ii) continuity axiom. $\mathcal{E}$ is called *bilinear functional energy* (BLFE) if, in addition, it satisfies the third axiom: (iii) contraction condition, i.e. $\forall u \in D(\mathcal{E})$, $u^* = u^+ \wedge 1 \in D(\mathcal{E})$ and $\mathcal{E}(u \pm u^*, u \mp u^*) \geq 0$.

For a the general theory of closed forms associated with Markov processes see (Ma, Rockner 1990).

Let $(\mathcal{L},D(\mathcal{L}))$ be the generator of a coercive form $(\mathcal{E},D(\mathcal{E}))$ on $L^2(X,\mu)$, i.e. the unique closed linear operator on $L^2(X,\mu)$ such that $1-\mathcal{L}$ is onto, $D(\mathcal{L}) \subset D(\mathcal{E})$ and $\mathcal{E}(u,v) =< -\mathcal{L}u,v >$ for all $u \in D(\mathcal{L})$ and $v \in D(\mathcal{E})$. Let $(T_t)_{t>0}$ be the strongly continuous contraction semigroup on $L^2(X,\mu)$ generated by $\mathcal{L}$ and $(G_\alpha)_{\alpha>0}$ the corresponding strongly continuous contraction semigroup (which exist according to the Hille-Yosida theorem).

A right process M with the state space X is *associated* with a BLFE $(\mathcal{E},D(\mathcal{E}))$ on $L^2(X,\mu)$ if the semigroup (P_t) of the process M is a μ-version[1] of the form semigroup (T_t). It has been proved (Albeverio, Ma, Rockner 1993) and (Ma, Rockner 1990) that only those BLFEs, which satisfy some regularity conditions can be associated with some right Markov processes and viceversa (Th.1.9 of (Albeverio, Ma, Rockner 1993)).

Prop. 4.2 from (Albeverio, Ma, Rockner 1993) states that two right Markov processes M and M' with state space X associated with a common quasi-regular BLFE $(\mathcal{E},D(\mathcal{E}))$ are stochastically equivalent (Ma, Rockner 1990). That means a quasi-regular BLFE

[1] I.e., for all $f \in L^2(X,\mu)$ the function $P_t f$ is a μ-version (differs on a set of μ-measure zero) of $T_t f$ for all $t > 0$.

characterizes a class of stochastically equivalent right Markov processes.

Let $M = (\Omega, \mathcal{F}, \mathcal{F}_t, x_t, P_x)$ be a right Markov process with the state space X. Now assume that X is a Lusin space (i.e. it is homeomorphic to a Borel subset of a compact metric space) and $\mathcal{B}(X)$ or $\mathcal{B}$ is its Borel σ-algebra. Assume also that μ is a σ-finite measure on $(X, \mathcal{B})$ and μ is a stationary measure of the process M. Let $X^{\#}$ another Lusin space (with $\mathcal{B}^{\#}$ its Borel σ-algebra) and $F : X \to X^{\#}$ be a measurable function. Let $\sigma(F)$ be the sub-σ-algebra of $\mathcal{B}$ generated by F. If μ is a probability measure then the projection operator between $L^2(X, \mathcal{B}, \mu)$ and $L^2(X, \sigma(F), \mu)$ is the conditional expectation $E_\mu[\cdot|F]$. Recall that E_μ is the expectation defined w.r.t. P_μ and that $P_\mu(A) = \int P_x(A) d\mu$, $A \in \mathcal{F}$. We denote by $\mu^{\#}$ the image of μ under F, i.e. $\mu^{\#}(A^{\#}) = \mu(F^{-1}(A^{\#}))$, for all $A^{\#} \in \mathcal{B}^{\#}$. In general, anything associated with $X^{\#}$ will carry the #-superscript symbol in this section.

Let $\mathcal{E}$ be the BLFE on $L^2(X, \mu)$ associated to M. F induces a form $\mathcal{E}^{\#}$ on $L^2(X^{\#}, \mu^{\#})$ by

$$\mathcal{E}^{\#}(u^{\#}, v^{\#}) = \mathcal{E}(u^{\#} \circ F, v^{\#} \circ F); \tag{3}$$

for $u^{\#}, v^{\#} \in D[\mathcal{E}^{\#}]$, where

$$D[\mathcal{E}^{\#}] = \{u^{\#} \in L^2(X^{\#}, \mu^{\#}) | u^{\#} \circ F \in D[\mathcal{E}]\}. \tag{4}$$

It can be shown (see Prop.1.4 in (Iscoe, McDonald 1990)), under a mild condition on the conditional expectation operator $E_\mu[\cdot|F]$ that $\mathcal{E}^{\#}$ is a BLFE. If, in addition, $\mathcal{E}^{\#}$ is quasi-regular then we can associate it a right Markov process $M^{\#} = (\Omega, \mathcal{F}, \mathcal{F}_t, x_t^{\#}, P_x^{\#})$ with the state space $X^{\#}$. The process $M^{\#}$ is called the *induced Markov process* w.r.t. to the proper map F. If the image of M under F is a right Markov process then $x_t^{\#} = F(x_t)$. The process $M^{\#}$ might have some different interpretations like a refinement of discrete transitions structure, or an approximation of continuous dynamics or an abstraction of the entire process. It is difficult to find a practical condition to impose on F, which would guarantee that $\mathcal{E}^{\#}$, as defined by (3) and (4), is also quasi-regular. To circumvent this problem, it is possible to restrict the original domain $D[\mathcal{E}^{\#}]$ and impose some regularity conditions on F (for more details, see (Iscoe, McDonald 1990)).

Assumption 1. *Suppose that $\mathcal{E}^{\#}$ is a quasi-regular BLFE.*

3 THE STOCHASTIC MODEL CHECKING PROBLEM

Let us consider $M = (\Omega, \mathcal{F}, \mathcal{F}_t, x_t, P_x)$ a strong Markov process, which is the semantics of a UCPS. For this strong Markov process we address a verification problem consisting of the *stochastic reachability problem* defined as follows. Given a set $A \in \mathcal{B}(X)$ and a time horizon $T > 0$, let us to define (Bujorianu 2004):

$$\begin{aligned} Reach_T(A) &= \{\omega \in \Omega \mid \exists t \in [0,T] : x_t(\omega) \in A\} \\ Reach_\infty(A) &= \{\omega \in \Omega \mid \exists t \geq 0 : x_t(\omega) \in A\}. \end{aligned} \tag{5}$$

These two sets are the sets of trajectories of M, which reach the set A (the flow that enters A) in the interval of time $[0,T]$ or $[0,\infty)$.

The reachability problem consists of determining the probabilities of such sets. The reachability problem is well-defined, i.e. $Reach_T(A)$, $Reach_\infty(A)$ are indeed measurable sets. Then the probabilities of reach events are

$$P(T_A < T) \text{ or } P(T_A < \infty) \tag{6}$$

where $T_A = \inf\{t > 0 | x_t \in A\}$ and P is a probability on the measurable space $(\Omega, \mathcal{F})$ of the elementary events associated to M. P can be chosen to be P_x (if we want to consider the trajectories, which start in x) or P_μ (if we want to consider the trajectories, which start in x_0 given by the distribution μ).

Usually a target set A in the state space is a level set for a given function $F : X \to \mathbb{R}$, i.e. $A = \{x \in X | F(x) > l\}$ (F can be chosen as the Euclidean norm or as the distance to the boundary of E). The probability of the set of trajectories, which hit A until time horizon $T > 0$ can be expressed as

$$P(\sup F(x_t) | t \in [0,T]) > l. \tag{7}$$

Our goal is to *define a new stochastic process $M^{\#}$ such that the probabilities (6) are preserved.*

Ideally, since (6) can be written as (7), $F(x_t)$ would represent the best candidate for defining a possible qualitative model reduction for M, which preserves the reach set probabilities. The main difficulty is that $F(x_t)$ is a Markov process only for special choices of F (Rogers, Pitman 1981). The problem is how to choose F well.

Note, if $A^{\#}$ is open in $X^{\#}$ and $A = F^{-1}(A^{\#})$, then we consider the two first hitting times T_A (w.r.t. M) and $T^{\#}_{A^{\#}}$ (w.r.t. $M^{\#}$) of A and $A^{\#}$, respectively. Recall that $T_A = \inf\{t > 0 | x_t \in A\}$.

The following results show that the stochastic model checking problem is solvable for uncertain cps.

Proposition 1. *Under the assumption.1, if μ is a probability measure and $\xi = +\infty$ (M has no killing), then*

$$E_\mu \exp(-T_A) \leq E_{\mu^{\#}} \exp(-T^{\#}_{A^{\#}}) \tag{8}$$

where E_μ (resp. $E_{\mu^{\#}}$) is the expectation defined w.r.t. P_μ (resp. $P_{\mu^{\#}}$).

If M is the semantics of a UCPS U, given a target state set $A \in \mathcal{B}(X)$, then the goal in the stochastic reachability analysis is to compute the probability $P_\mu(T_A \leq T)$ for a finite horizon time $T > 0$. We now translate the relation (8) in terms of probability of the reachable sets.

Proposition 2. *Under the assumption.1, if μ is a probability measure, then*

$$P_\mu(T_A \quad \leq \quad T) \leq eK \min\{T\mathcal{E}^{\#}(u^{\#},u^{\#}) + \tag{9}$$

$$< \quad u^{\#},u^{\#} >_{\mu^{\#}} |u^{\#} \in D(\mathcal{E}^{\#}),\, u^{\#} \geq 1, \tag{10}$$

$$\mu^{\#} - a.e.\ on\ A^{\#}\} \tag{11}$$

where $K > 0$ is the sector constant of $\mathcal{E}$.

4 ADAPTING VERIFICATION TO CO-EVOLUTION

The idea is to apply a “state space reduction” technique based on the general ‘induced BLFEs’ method to achieve qualitative model reductions for Ucps. With this technique, the semantics of Ucps are ‘approximated’ by a one-dimensional stochastic process with a much smaller state space.

4.1 Qualitative Model Reduction

The stochastic reachability definition gives the idea to introduce the following concept of qualitative model reduction for Ucps.

Definition 1. *Given a right Markov process M defined on the Lusin state space $(X,\mathcal{B})$, and $F : X \to \mathbb{R}$ a measurable weight function, suppose that assumption.1 is fulfilled. The process $M^{\#}$ associated to the induced BLFE $\mathcal{E}^{\#}$ under function F is called a qualitative model reduction of M.*

Let U be a UCPS and M its semantics. Suppose that M is a right Markov process defined on the Lusin state space $(X,\mathcal{B})$.

Definition 2. *Any UCPS $U^{\#}$ whose semantics is a qualitative model reduction of M is called a qualitative model reduction of U.*

Let U be a Ucps and M its semantics (that is a right Markov process, with the state space X).

Proposition 3. *If M is a diffusion then any qualitative model reduction $M^{\#}$ of M is a diffusion.*

Proposition 4. *If M is a jump process then any qualitative model reduction $M^{\#}$ of M is again a jump process.*

Proof. This statement can be obtained as a consequence of the abstract version of the *Kolmogorov backward equations* (Davis 1993)

$$\frac{\partial}{\partial t} P_t f(x) = L P_t f(x),\, P_0 f = f,\, f \in D(\mathcal{L}) \tag{12}$$

and the equality (14). If the equations (12) are associated to an initial diffusion process (resp. jump process) then the relation (14) allow to obtain the fact that the transition probabilities of the induced process satisfy a similar equation, such that the induced process is still a diffusion process (resp. jump process). The same conclusion can be obtain using the stochastic calculus of BLFEs (Iscoe, McDonald 1990).■

Since the semantics of a Ucps is, in most cases, a stochastic process, which can be viewed an interleaving between some diffusion processes and a jump process (see (Bujorianu, Lygeros 2004) for a very general model for Ucps and its semantics as a Markov string), we can write the following result as a corollary of Prop.3.

Proposition 5. *Any qualitative model reduction of a Ucps is again a Ucps.*

Let $(\mathcal{L},D(\mathcal{L}))$ and $(\mathcal{L}^{\#},D(\mathcal{L}^{\#}))$ be the generators of $\mathcal{E}$ and $\mathcal{E}^{\#}$, respectively. For the following results we suppose that the Ass.1 is fulfilled.

Proposition 6. *The generators $\mathcal{L}$ and $\mathcal{L}^{\#}$ are related as follows*

$$\mathcal{L}(u^{\#} \circ F) = \mathcal{L}^{\#} u^{\#} \circ F, \forall u^{\#} \in D(\mathcal{L}^{\#}) \tag{13}$$

Theorem 7. *For all $A^{\#} \in \mathcal{B}^{\#}(X^{\#})$ and for all $t > 0$ we have*

$$p_t^{\#}(Fx, A^{\#}) = p_t(x, F^{-1}(A^{\#})) \tag{14}$$

where $(p_t^{\#})$ and (p_t) are the transition functions of $M^{\#}$ and M, respectively.

Proof. Let $F^{\#}$ be defined as $F^{\#} : \mathcal{B}^b(X^{\#}) \to \mathcal{B}^b(X)$; $F^{\#}u^{\#} = u^{\#} \circ F$. Then (13) becomes $(\mathcal{L} \circ F^{\#})u^{\#} = (F^{\#} \circ \mathcal{L}^{\#})u^{\#}, \forall u^{\#} \in D(\mathcal{L}^{\#})$ $(**)$. For a strong Markov process, the opus of the kernel operator is the inverse operator of the infinitesimal generator of the process. Now, from $(**)$ we get a similar relation between the kernel operators V and $V^{\#}$ of the processes M and $M^{\#}$, i.e. $F^{\#} \circ V^{\#} = V \circ F^{\#}$ on $\mathcal{B}^b(X^{\#})$, or

$$V^{\#} u^{\#} \circ F = V(u^{\#} \circ F), \forall u^{\#} \in \mathcal{B}^b(X^{\#}) \tag{15}$$

since if $u^{\#} \in \mathcal{B}^b(X^{\#})$ then $V^{\#}u^{\#} \in D(\mathcal{L}^{\#})$. For $u^{\#} = 1_{A^{\#}}$ (the indicator function of $A^{\#}$), by the kernel operator integral definition, we obtain (14).■

Relation (15) implies the following corollary:

Corollary 8. *The semigroups $(P_t^{\#})$ and (P_t) of $M^{\#}$ and M are related by*

$$P_t^{\#} u^{\#} \circ F = P_t(u^{\#} \circ F), \forall u^{\#} \in \mathcal{B}^b(X^{\#}). \tag{16}$$

4.2 Adaptive Bisimulation

In this subsection we define a new concept of adaptive bisimulation for cps. This concept is defined as measurable relation, which induces equivalent BLFEs on the quotient spaces. In defining adaptive bisimulation, we do not impose the equivalence of the quotient processes, which might not have Markovian properties (Rogers, Pitman 1981), but we impose the equivalence of the qualitative model reductions (that can differ from the quotient processes) associated with the induced BLFEs, with respect to the projection maps.

Let $(X, \mathcal{B}(X))$ and $(Y, \mathcal{B}(Y))$ be Lusin spaces and let $\mathcal{R} \subset X \times Y$ be a relation such that $\Pi^1(\mathcal{R}) = X$ and $\Pi^2(\mathcal{R}) = Y$. We define the equivalence relation on X that is induced by the relation $\mathcal{R} \subset X \times Y$, as the transitive closure of $\{(x,x')|\exists y \text{ s.t. } (x,y) \in \mathcal{R} \text{ and } (x',y) \in \mathcal{R}\}$. Analogously, the induced (by $\mathcal{R}$) equivalence relation on Y can be defined. We write $X/_{\mathcal{R}}$ and $Y/_{\mathcal{R}}$ for the sets of equivalence classes of X and Y induced by $\mathcal{R}$. We denote the equivalence class of $x \in X$ by $[x]$. Let

$$\mathcal{B}^{\#}(X) = \mathcal{B}(X) \cap \{A \subset X| \text{ if } x \in A \text{ and } [x] = [x'] \text{ then } x' \in A\}$$

be the collection of all Borel sets, in which any equivalence class of X is either totally contained or totally not contained. It can be checked that $\mathcal{B}^{\#}(X)$ is a σ-algebra. Let $\pi_X : X \to X/_{\mathcal{R}}$ be the mapping that maps each $x \in X$ to its equivalence class and let

$$\mathcal{B}(X/_{\mathcal{R}}) = \{A \subset X/_{\mathcal{R}} | \pi_X^{-1}(A) \in \mathcal{B}^{\#}(X)\}.$$

Then $(X/_{\mathcal{R}}, \mathcal{B}(X/_{\mathcal{R}}))$, which is a measurable space, is called the quotient space of X w.r.t. $\mathcal{R}$. The quotient space of Y w.r.t. $\mathcal{R}$ is defined in a similar way. We define a bijective mapping $\psi : X/_{\mathcal{R}} \to Y/_{\mathcal{R}}$ as

$$\psi([x]) = [y] \text{ if } (x,y) \in \mathcal{R} \text{ for some } x \in [x] \text{ and some } y \in [y].$$

We say that the relation $\mathcal{R}$ is *measurable* if X and Y if for all $A \in \mathcal{B}(X/_{\mathcal{R}})$ we have $\psi(A) \in \mathcal{B}(Y/_{\mathcal{R}})$ and vice versa, i.e. ψ is a homeomorphism. Then the real measurable functions defined on $X/_{\mathcal{R}}$ can be identified with those defined on $Y/_{\mathcal{R}}$ through the homeomorphism ψ. We can write $\mathcal{B}^b(X/_{\mathcal{R}}) \overset{\psi}{\cong} \mathcal{B}^b(Y/_{\mathcal{R}})$. Moreover, these functions can be thought of as real functions defined on X or Y measurable w.r.t. $\mathcal{B}^{\#}(X)$ or $\mathcal{B}^{\#}(Y)$.

Assumption 2. *Suppose that $X/_{\mathcal{R}}$ and $Y/_{\mathcal{R}}$ with the topologies induced by projection mappings are Lusin spaces.*

Suppose we have given two right Markov processes M and W with the state spaces X and Y. Assume that μ (resp. ν) is a stationary measure of the process M (resp. W). Let $\mu/_{\mathcal{R}}$ (resp. $\nu/_{\mathcal{R}}$) the image of μ (resp. ν) under π_X (resp. π_Y). Let $\mathcal{E}$ (resp. $\mathcal{F}$) the quasi-regular BLFE corresponding to M (resp. W). The equivalence between the induced processes can be used to define a new bisimulation between Markov processes, as follows.

Definition 3. *Under assumptions 1 and 2, a measurable relation $\mathcal{R} \subset X \times Y$ is a bisimulation between M and W if the mappings π_X and π_Y define the same induced BLFE on $L^2(X/_{\mathcal{R}}, \mu/_{\mathcal{R}})$ and $L^2(Y/_{\mathcal{R}}, \nu/_{\mathcal{R}})$, respectively.*

This definition states that M and W are bisimilar if $\mathcal{E}/_{\mathcal{R}} = \mathcal{F}/_{\mathcal{R}}$. Here, $\mathcal{E}/_{\mathcal{R}}$ (resp. $\mathcal{F}/_{\mathcal{R}}$) is the induced BLFE of $\mathcal{E}$ (resp. $\mathcal{F}$) under the mapping π_X (resp. π_Y). Clearly, this can be possible iff $\mu/_{\mathcal{R}} = \nu/_{\mathcal{R}}$.

Assumption 3. *Suppose that $\mathcal{E}/_{\mathcal{R}}$ and $\mathcal{F}/_{\mathcal{R}}$ are quasi-regular BLFE.*

Denote the Markov process associated to $\mathcal{E}/_{\mathcal{R}}$ (resp. $\mathcal{F}/_{\mathcal{R}}$) by $M/_{\mathcal{R}}$ (resp. $W/_{\mathcal{R}}$).

Proposition 9. *Under assumptions 1, 2 and 3, M and W are stochastic bisimilar under $\mathcal{R}$ iff their qualitative model reductions $M/_{\mathcal{R}}$ and $W/_{\mathcal{R}}$ with respect to π_X and, respectively π_Y are $\mu/_{\mathcal{R}}$-equivalent.*

Let U and U' be two UCPSs, with the semantics M and W, strong Markov processes defined on the state spaces $(X, \mathcal{B}(X))$ and $(Y, \mathcal{B}(Y))$, respectively.

Definition 4. *U and U' are called bisimilar if there exist a bisimulation relation under which their semantics M and W are bisimilar*

4.3 An Example

Let us recall the chemically reacting system case study from (Singh, Hespanha 2005), where it is investigated using the theory of polynomial stochastic hybrid systems. Consider a system of n species X_j, $j = 1,..,n$, inside a fixed volume V involved in K reactions of the form

$$(R_i) u_{ai}A_i + u_{i1}X_1 + \ldots + u_{in}X_n \xrightarrow{c_i} \nu_{i1}X_1 + \ldots + \nu_{in}X_n, \forall i \in \{1, \ldots, K\}$$

where the species A_i have a constant number of molecules. The meaning and the assumptions about the coefficients of the reaction equation are given in (Singh, Hespanha 2005). c_i is a reaction parameter which is used in defining the probability that a particular reaction takes place on $(t, t+dt)$. The system is characterized by the trivial dynamics $\dot{x}=0$, $x=[x_1, x_2, \ldots, x_n]^T$, a family of K *reset maps* $x=\phi_i(x^-)$, $\phi_i : \mathbb{R}^n \to \mathbb{R}^n$, and a corresponding family of *transition intensities* $\lambda_i : \mathbb{R}^n \to [0,\infty)$, $\forall i = 1,..,K$. For each $i=1,..,K$, the reset map ϕ_i and the corresponding λ_i is uniquely defined by the i^{th} reaction equation and given by $x \mapsto \phi_i(x)$, $\phi_i(x) = x + [\nu_{i1} - u_{i1}, \nu_{i2} - u_{i2}, \ldots, \nu_{in} - u_{in}]^T$; $\lambda_i(x) = c_i h_i(x)$, where U_i represents the number of distinct molecular reactant combinations present in V at time t for the reaction R_i. The executions of such a system are defined in (Singh, Hespanha 2005).

Now we apply the method of qualitative model reduction to this process. We can show that executions of this cps form a particular kind of right Markov process called jump process (Davis 1993). The extended generator (Th.1 (Singh, Hespanha 2005)) is $(L\psi)(x) = \sum_{i=1}^{K} (\psi(\phi_i(x)) - \psi(x))\lambda_i(x)$, $\psi \in D(L)$.

Let us consider a proper map $F : \mathbb{R}^n \to \mathbb{R}$ and write the generator of the induced process for
$\psi^{\#} \circ F$, $\psi^{\#} \in D(L^{\#})$:
$L(\psi^{\#} \circ F)(x) = \sum_{i=1}^{K} (\psi^{\#}(F(\phi_i(x))) - \psi^{\#}(F(x)))\lambda_i(x)$

Define $\phi_i^{\#} : \mathrm{Im}F \to \mathbb{R}$ by $\phi_i^{\#}(Fx) = F(\phi_i(x))$ and $\lambda_i^{\#} : \mathrm{Im}F \to \mathbb{R}$ by $\lambda_i^{\#}(Fx) = \lambda_i(x)$. In order to have these two function well-defined we need to impose some compatibility conditions between F and reset maps ϕ_i and their corresponding transition intensities λ_i as follows: $Fx = Fx' \Rightarrow F(\phi_i(x)) = F(\phi_i(x'))$ and $\lambda_i(x) = \lambda_i(x')$. This means that F preserves the jumps (reset maps and transition intensities), i.e. the pre-jump locations have the same image under F then the intensities of transition should be equal and the post-jump locations have the same image under F. Using (13), the generator of the induced process is

$$\begin{aligned} L^{\#}\psi^{\#}(x^{\#}) &= \sum_{i=1}^{K} (\psi^{\#}(\phi_i^{\#}(x^{\#})) - \psi^{\#}(x^{\#}))\lambda_i^{\#}(x^{\#}); \\ & \quad x^{\#} = Fx;\ x \in X. \end{aligned}$$

For simplicity, we suppose that the reactions R_i are reversible in time. Then the generator is self-adjoint (or Hermitian). The (symmetric) quasi-regular energy bilinear form on $L^2(\mathbb{R}^n, \mu)$ associated to the given process (with μ a stationary distribution) can be written

$\mathcal{E}(\psi, \varphi) = \sum_{i=1}^{K} \int_{\mathbb{R}^n} (\psi(\phi_i(x)) - \psi(x))(\varphi(\phi_i(x)) - \varphi(x))\lambda_i(x)\mu(dx)$

Then the induced energy bilinear form $\mathcal{E}^{\#}$ on $L^2(\mathbb{R}, \mu^{\#})$ (where $\mu^{\#}$ is the image of μ under F) w.r.t. F is

$$\begin{aligned} \mathcal{E}^{\#}(\psi^{\#}, \varphi^{\#}) &= \sum_{i=1}^{K} \int_{\mathbb{R}^n} [\psi^{\#}(\phi_i^{\#}(Fx)) - \psi^{\#}(Fx)] \\ & \quad [\varphi^{\#}(\phi_i^{\#}(Fx)) - \varphi^{\#}(Fx)]\lambda_i^{\#}(Fx) \\ & \quad \mu(dx) \\ &= \sum_{i=1}^{K} \int_{\mathbb{R}} [\psi^{\#}(\phi_i^{\#}(x^{\#})) - \psi^{\#}(x^{\#})][\varphi^{\#} \\ & \quad (\phi_i^{\#}(x^{\#})) - \varphi^{\#}(x^{\#})]\lambda_i^{\#}(x^{\#})\mu^{\#}(dx^{\#}). \end{aligned}$$

Clearly, $\mathcal{E}^{\#}$ is associated to a jump process - thequalitative model reduction of the given process. In this particular case, the induced process is exactly the image under F of the initial jump process.

5 CONCLUSIONS

In this paper, we have used the concept of energy, which is a key ingredient of Hilbertean formal methods, to define qua;itative model reduction and behavioral equivalence for cyber-physical systems operating in random environments. Energy is a versatile analytical concept that characterizes in a subtle way the interaction between computation and physics, as well as their co-evolution.

Adaptive bisimulation means the energy preservation of the stochastic processes generated by the cyber-physical system evolutions. The energy concept can be also used to define qualitative model reductions for cyber-physical systems. Given an qualitative model reduction function that reduces the state space, we have defined a standard construction that associates a qualitative model reduction (called standard) on the reduced state space. The mathematical results from Section 4.1 show that the qualitative model reduction method preserves important analytic properties (related to HFM). Two uncertain CPS are adaptive bisimilar if they have the same energy. The theorem from Section 4.2 shows that two uncertain CPS are adaptive bisimilar iff their standard qualitative model reductions are equivalent as Markov processes.

We have formulated the stochastic model checking problem (a subproblem of stochastic reachability analysis, corresponding to the probabilistic model checking of Markov chains). We proved two results that show that the problem is solvable for uncertain cyber-physical systems. The mathematical results from Section 3 provide a upper bound for the reach set probabilities. In this way, one can prove that the probability of reaching a state in a certain set can be small enough.

The most closely related model is that of *stochastic hybrid automata* (Bujorianu 2004). These automata are not necessarily embedded systems and their hybrid behavior is often an internal feature (as for cars, aircraft, mobile robots and so on) rather than the interaction with a physical environment (a feature of embedded systems). Cyber-physical systems are also networked.

In following work we will refine the formal framework presented in this paper to be used for nanoscience.

ACKNOWLEDGEMENTS

This work was funded by the EPSRC project EP/E050441/1 CICADA.

REFERENCES

Accardi L., Ohya M., Watanabe N., 2006. *Quantum Information and Computing* World Scientific.

Albeverio, S., Ma, Z.M., Rockner, M., 1993. Quasi-regular Dirichlet Forms and Markov Processes. *J. of Functional Analysis* 111: 118-154.

Bujorianu, M.C., Bujorianu M.L., 2007a. Towards Hilbertean Formal Methods *Proc. of the 7th International Conference on Application of Concurrency to System Design ACSD* IEEE Press.

Bujorianu, M.C., Bujorianu, M.L., 2007b. An integrated specification framework for embedded systems, *Proc. of SEFM*, IEEE Press.

Bujorianu, M.C., Bujorianu M.L., 2008a. A Randomized Model for Communicating Embedded Systems. *Proceedings of the 16th Mediterranean Conference on Control and Automation*, IEEE Press.

Bujorianu, M.L., Bujorianu, M.C., 2008b. Bisimulation, Logic and Mobility for Markovian Systems, In: *Proc of 18th International Symposium on Mathematical Theory of Networks and Systems (MTNS08)*, SIAM.

Bujorianu, M.L., Bujorianu, M.C., Blom H., 2008. Approximate Abstractions of Stochastic Hybrid Systems, *Proc. of the 17th IFAC World Congress*, Elsevier.

Bujorianu, M.L., Katoen J., 2008. Symmetry reduction for stochastic hybrid systems. In: *Proc. of IEEE 47th Conference on Decision and Control*, IEEE press.

Bujorianu, M.L., Bujorianu, M.C. 2006. A Model Checking Strategy for a Performance Measure of Fluid Stochastic Models, In: *European Performance Engineering Workshop (EPEW)*, Springer LNCS 4054, pp. 93-107.

Bujorianu, M.L., Lygeros, J., 2004. General Stochastic Hybrid Systems: Modelling and Optimal Control. *Proc. 43th Conference in Decision and Control*, IEEE Press: 182-187.

Bujorianu, M.L. 2004. Extended Stochastic Hybrid Systems and their Reachability Problem. In *Hybrid Systems: Computation and Control*, Springer LNCS 2993: 234-249.

Bujorianu, M.L., Lygeros, J., Bujorianu, M.C., 2005a. Abstractions of Stochastic Hybrid System. *Proc. 44th Conference in Decision and Control*. IEEE Press.

Bujorianu, M.L., Lygeros, J., Bujorianu, M.C., 2005b. Bisimulation for General Stochastic Hybrid Systems. In *Proc. Hybrid Systems: Computation and Control*, Springer LNCS 3414: 198-216.

Davis, M.H.A. 1993. *Markov Models and Optimization* Chapman & Hall.

Ethier, S.N., Kurtz, T.G., 1986. *Markov Processes: Characterization and Convergence*. John Wiley and Sons.

Haghverdi, E., Tabuada, P., Pappas, G.J., 2005. Bisimulation Relations for Dynamical, Control and Hybrid Systems. *Theor. Comput. Science*, 342(2-3):229-261.

Hornyak, G., Dutta, J., Tibbals H.J., Rao A.K. 2008. *Introduction to Nanoscience* CRC Press.

Iscoe, I., McDonald, D., 1990. Induced Dirichlet Forms and Capacitary Inequalities. *Ann. Prob.* 18 (3): 1195-1221.

Ma, M., Rockner, M., 1990. *The Theory of (Non-Symmetric) Dirichlet Forms and Markov Processes* Springer Verlag.

Rogers, L.C.G., Pitman, J.W., 1981. Markov Functions. *Ann. Prob.*, 9 (4): 573-582.

Singh, A., Hespanha, J.P., 2005. Models for Multi-Specie Chemical Reactions Using Polynomial Stochastic Hybrid Systems. *Proc. of 44th Conference in Decision and Control*, IEEE Press.

Tabuada P. 2006. *Cyber-Physical Systems: Position Paper* presented at NSF Workshop on Cyber-Physical Systems.

SHORT PAPERS

SUBOPTIMAL DUAL CONTROL ALGORITHMS FOR DISCRETE-TIME STOCHASTIC SYSTEMS UNDER INPUT CONSTRAINT

Andrzej Krolikowski
Poznan University of Technology, Institute of Control and Information Engineering, Poland
Andrzej.Krolikowski@put.poznan.pl

Dariusz Horla
Poznan University of Technology, Institute of Control and Information Engineering, Poland
Dariusz.Horla@put.poznan.pl

Keywords: Input constraint, Suboptimal dual control.

Abstract: The paper considers a suboptimal solution to the dual control problem for discrete-time stochastic systems under the amplitude-constrained control signal. The objective of the control is to minimize the two-step quadratic cost function for the problem of tracking the given reference sequence. The presented approach is based on the MIDC (Modified Innovation Dual Controller) derived from an IDC (Innovation Dual Controller) and the TSDSC (Two-stage Dual Suboptimal Control. As a result, a new algorithm, i.e. the two-stage innovation dual control (TSIDC) algorithm is proposed. The standard Kalman filter equations are applied for estimation of the unknown system parameters. Example of second order system is simulated in order to compare the performance of proposed control algorithms. Conclusions yielded from simulation study are given.

1 INTRODUCTION

The problem of the optimal control of stochastic systems with uncertain parameters is inherently related with the dual control problem where the learning and control processes should be considered simultaneously in order to minimize the cost function. In general, learning and controlling have contradictory goals, particularly for the finite horizon control problems. The concept of duality has inspired the development of many control techniques which involve the dual effect of the control signal. They can be separated in two classes: explicit dual and implicit dual (Bayard and Eslami, 1985). Unfortunately, the dual approach does not result in computationally feasible optimal algorithms. A variety of suboptimal solutions has been proposed, for example: the innovation dual controller (IDC) (R. Milito and Cadorin, 1982) and its modification (MIDC) (Królikowski and Horla, 2007), the two-stage dual suboptimal controller (TSDSC) (Maitelli and Yoneyama, 1994) or the pole-placement (PP) dual control (N.M. Filatov and Keuchel, 1993).

Other controllers like minimax controllers (Sebald, 1979), Bayes controllers (Sworder, 1966), MRAC (Model Reference Adaptive Controller) (Åström and Wittenmark, 1989), LQG controller where unknown system parameters belong to a finite set (D. Li and Fu,) or Iteration in Policy Space (IPS) (Bayard, 1991) are also possible.

The IPS algorithm and its reduced complexity version were proposed by Bayard (Bayard, 1991) for a general nonlinear system. In this algorithm the stochastic dynamic programming equations are solved forward in time ,using a nested stochastic approximation technique. The method is based on a specific computational architecture denoted as a H block. The method needs a filter propagating the state and parameter estimates with associated covariance matrices.

In (Królikowski, 2000), some modifications including input constraint have been introduced into the original version of the IPS algorithm and its performance has been compared with MIDC algorithm.

In this paper, a new algorithm, i.e. the two-stage innovation dual control (TSIDC) algorithm is proposed which is the combination of the IDC approach and the TSDSC approach. Additionally, the amplitude constraint of control input is taken into consideration for algorithm derivation.

Performance of the considered algorithms is illustrated by simulation study of second-order system with control signal constrained in amplitude.

2 CONTROL PROBLEM FORMULATION

Consider a discrete-time linear single-input single-output system described by ARX model

$$A(q^{-1})y_k = B(q^{-1})u_k + w_k, \tag{1}$$

where $A(q^{-1}) = 1 + a_{1,k}q^{-1} + \cdots + a_{na,k}q^{-na}$, $B(q^{-1}) = b_{1,k}q^{-1} + \cdots + b_{nb,k}q^{-nb}$, y_k is the output available for measurement, u_k is the control signal, $\{w_k\}$ is a sequence of independent identically distributed gaussian variables with zero mean and variance σ_w^2. Process noise w_k is statistically independent of the initial condition y_0. The system (1) is parametrized by a vector θ_k containing $na + nb$ unknown parameters $\{a_{i,k}\}$ and $\{b_{i,k}\}$ which in general can be assumed to vary according to the equation

$$\underline{\theta}_{k+1} = \Phi\underline{\theta}_k + \underline{e}_k \tag{2}$$

where Φ is a known matrix and $\{\underline{e}_k\}$ is a sequence of independent identically distributed gaussian vector variables with zero mean and variance matrix R_e. Particularly, for the constant parameters we have

$$\underline{\theta}_{k+1} = \underline{\theta}_k = \underline{\theta} = (b_1, \cdots, b_{nb}, a_1, \cdots a_{na})^T, \tag{3}$$

and then $\Phi = I$, $\underline{e}_k = 0$ in (2).

The control signal is subjected to an amplitude constraint

$$\mid u_k \mid \leq \alpha \tag{4}$$

and the information state I_k at time k is defined by

$$I_k = [y_k, ..., y_1, u_{k-1}, ..., u_0, I_0] \tag{5}$$

where I_0 denotes the initial conditions.

An admissible control policy Π is defined by a sequence of controls $\Pi = [u_0, ..., u_{N-1}]$ where each control u_k is a function of I_k and satisfies the constraint (4). The control objective is to find a control policy Π which minimizes the following expected cost function

$$J = E[\sum_{k=0}^{N-1}(y_{k+1} - r_{k+1})^2] \tag{6}$$

where $\{r_k\}$ is a given reference sequence. An admissible control policy minimizing (6) can be labelled by CCLO (Constrained Closed-Loop Optimal) in keeping with the standard nomenclature, i.e. $\Pi^{\text{CCLO}} = [u_0^{\text{CCLO}}, ..., u_{N-1}^{\text{CCLO}}]$. This control policy has no closed form, and control policies presented in the following section can be viewed as a suboptimal approach to the Π^{CCLO}.

3 SUBOPTIMAL DUAL CONTROL ALGORITHMS

In this section, we shall briefly describe algorithms being an approximate solution to the problem formulated in Section 2. To this end, the method for estimation of system parameters $\underline{\theta}_k$ is needed.

3.1 Estimation Method

The system (1) can be expressed as

$$y_{k+1} = \underline{s}_k^T \underline{\theta}_{k+1} + w_{k+1} \tag{7}$$

where

$$\begin{aligned} \underline{s}_k &= (u_k, u_{k-1}, \ldots, u_{k-nb+1}, -y_k, \ldots, -y_{k-na+1})^T = \\ &= (u_k, \underline{s}_k^{*^T})^T. \end{aligned} \tag{8}$$

The estimates $\hat{\underline{\theta}}_k$ needed to implement dual control algorithms can be obtained in many ways. A common way is to use the standard Kalman filter in a form of suitable recursive procedure for parameter estimation, i.e.

$$\begin{aligned} \hat{\underline{\theta}}_{k+1} &= \Phi\hat{\underline{\theta}}_k + \underline{k}_k\varepsilon_k & (9)\\ \underline{k}_k &= \Phi P_k \underline{s}_{k-1}[\underline{s}_{k-1}^T P_k \underline{s}_{k-1} + \sigma_w^2]^{-1} & (10)\\ P_{k+1} &= [\Phi - \underline{k}_k \underline{s}_{k-1}^T]P_k\Phi^T + R_e, & (11)\\ \varepsilon_k &= y_k - \underline{s}_{k-1}^T \hat{\underline{\theta}}_k, & (12) \end{aligned}$$

where ε_{k+1} is the innovation which will be used later on to construct the suboptimal dual control algorithm.

The following partitioning is introduced for parameter covariance matrix P_k

$$P_k = \begin{bmatrix} p_{b_1,k} & \underline{p}_{b_1\underline{\theta}^*,k}^T \\ \underline{p}_{b_1\underline{\theta}^*,k} & P_{\underline{\theta}^*,k} \end{bmatrix} \tag{13}$$

corresponding to the partition of $\underline{\theta}_k$

$$\underline{\theta}_k = (b_{1,k}, \underline{\theta}_k^{*T})^T \tag{14}$$

with

$$\underline{\theta}_k^* = (b_{2,k}, \ldots, b_{nb,k}, a_{1,k}, \ldots, a_{na,k})^T. \tag{15}$$

3.2 Two-stage Dual Suboptimal Control (TSDSC) Algorithm

The TSDSC method proposed in (Maitelli and Yoneyama, 1994) has been derived for system (1) with stochastic parameters (2). Below this method is extended for the input-constrained case. The cost function considered for TSDSC is given by

$$J = \frac{1}{2}E[(y_{k+1} - r)^2 + (y_{k+2} - r)^2 | I_k] \tag{16}$$

and according to (Maitelli and Yoneyama, 1994) can be obtained as a quadratic form in u_k and u_{k+1}, i.e.

$$J = \frac{1}{2}[au_k + bu_{k+1} + cu_k u_{k+1} + du_k^2 + eu_{k+1}^2] \quad (17)$$

where a,b,c,d,e are expressions depending on current data $\underline{s}_k^*$, reference signal r and parameter estimates $\hat{\underline{\theta}}_k$ (Maitelli and Yoneyama, 1994). Solving a necessary optimality condition the unconstrained control signal is

$$u_k^{\text{TSDSC,un}} = \frac{bc - 2ae}{4de - c^2}. \quad (18)$$

This control law has been taken for simulation analysis in (Maitelli and Yoneyama, 1994). Imposing the cutoff the constrained control signal is

$$u_k^{\text{TSDSC,co}} = sat(u_k^{\text{TSDSC,un}}; \alpha). \quad (19)$$

The cost function (27) can be represented as a quadratic form

$$J = \frac{1}{2}[\underline{u}_k^T A \underline{u}_k + \underline{b}^T \underline{u}_k] \quad (20)$$

where $\underline{u}_k = (u_k, u_{k+1})^T$, and

$$A = \begin{bmatrix} d & \frac{1}{2}c \\ \frac{1}{2}c & e \end{bmatrix}, \underline{b} = \begin{bmatrix} a \\ b \end{bmatrix}. \quad (21)$$

The condition $4de - c^2 > 0$ together with $d > 0$ implies positive definitness and guarantees convexity. Minimization of (30) under constraint (4) is a standard QP problem resulting in $\underline{u}_k^{\text{TSDSC,qp}}$. The constrained control $u_k^{\text{TSDSC,qp}}$ is then applied to the system in receding horizon framework.

3.3 Two-stage Innovation Dual Suboptimal Control (TSIDSC) Algorithm

A modified version of the TSDSC algorithm is given below where innovation term is included to the cost function

$$J = \frac{1}{2}E[(y_{k+1} - r)^2 + (y_{k+2} - r)^2 - \lambda_{k+1}\varepsilon_{k+1}^2 | I_k] \quad (22)$$

where $\lambda_{k+1} \geq 0$ is the learning weight, and ε_{k+1} is the innovation, see (16). Incorporating the term $-\lambda_{k+1}\varepsilon_{k+1}^2$ in the cost function makes the parameter estimation process to accelerate and consequently to improve the future control performance. Taking (2) and (7) into account it can be seen that

$$\varepsilon_{k+1} = \underline{s}_k^T[\Phi(\underline{\theta}_k - \hat{\underline{\theta}}_k) + (\Phi - I)\hat{\underline{\theta}}_k] + \underline{s}_k^T \underline{e}_k + w_{k+1}, \quad (23)$$

and consequently

$$\begin{aligned} E[\varepsilon_{k+1}^2 | I_k] &= \underline{s}_k^T \Phi P_k \Phi^T \underline{s}_k + \underline{s}_k^T (\Phi - I)\hat{\underline{\theta}}_k \hat{\underline{\theta}}_k^T (\Phi - I)^T + \\ &+ \underline{s}_k^T R_e \underline{s}_k + \sigma_w^2 = \\ &= \underline{s}_k^T [\Phi P_k \Phi^T + (\Phi - I)\hat{\underline{\theta}}_k \hat{\underline{\theta}}_k^T (\Phi - I)^T + R_e]\underline{s}_k + \sigma_w^2 = \\ &= \underline{s}_k^T \Sigma_k \underline{s}_k + \sigma_w^2. \quad (24) \end{aligned}$$

Introducing the partitioning for matrix Σ_k

$$\Sigma_k = \begin{bmatrix} \sigma_{11,k} & \underline{\sigma}_{1,k}^T \\ \underline{\sigma}_{1,k} & \Sigma_k^* \end{bmatrix}. \quad (25)$$

Keeping (8) in mind we have

$$E[\varepsilon_{k+1}^2 | I_k] = fu_k^2 + gu_k + h, \quad (26)$$

where $f = \sigma_{11,k}$, $g = \underline{\sigma}_{1,k}^T \underline{s}_k^*$, $h = \underline{s}_k^{*T} \Sigma_k{}^* \underline{s}_k^* + \sigma_w^2$ are expressions known at the current time instant k.

Finally, the cost function J including terms depending only on u_k and u_{k+1} takes the form

$$\begin{aligned} J = \frac{1}{2}[&au_k + bu_{k+1} + cu_k u_{k+1} + du_k^2 + eu_{k+1}^2 - \\ &-\lambda_{k+1}(fu_k^2 + gu_k)] \quad (27) \end{aligned}$$

Solving a necessary optimality condition the unconstrained control signal is

$$u_k^{\text{TSIDSC,un}} = \frac{bc - 2ae - 2eg}{4de - c^2 - 4ef\lambda_{k+1}}. \quad (28)$$

Imposing the cutoff the constrained control signal is

$$u_k^{\text{TSIDSC,co}} = sat(u_k^{\text{TSIDSC,un}}; \alpha). \quad (29)$$

The cost function (27) can again be represented as a quadratic form

$$J = \frac{1}{2}[\underline{u}_k^T A \underline{u}_k + \underline{b}^T \underline{u}_k] \quad (30)$$

where $\underline{u}_k = (u_k, u_{k+1})^T$, and correspondingly to (21)

$$A = \begin{bmatrix} d - \lambda_{k+1}f & \frac{1}{2}c \\ \frac{1}{2}c & e \end{bmatrix}, \underline{b} = \begin{bmatrix} a - \lambda_{k+1}g \\ b \end{bmatrix}. \quad (31)$$

The weight λ_{k+1} has influence on positive definitness of the quadratic form. Minimization of (30) under constraint (4) is again the QP problem resulting in $\underline{u}_k^{\text{TSIDSC,qp}}$. The constrained control $u_k^{\text{TSIDSC,qp}}$ is then applied to the system in receding horizon framework.

4 SIMULATION TESTS

Performance of the described control methods is illustrated through the example of a second-order system with the following true values: $a_1 = -1.8$, $a_2 = 0.9$, $b_1 = 1.0$, $b_2 = 0.5$, where the Kalman filter algorithm

(9)-(12) was applied for estimation. The initial parameter estimates were taken half their true values with $P_0 = 10I$. The reference signal was a square wave ±3, and then the minimal value of constraint α ensuring the tracking is $\alpha_{min} = 3\frac{|A(1)|}{|B(1)|} = 0.2$. Fig. 1 shows the reference, output and input signals during tracking process under the constraint $\alpha = 1$ for both TSDSC and TSIDSC control policies. The controls $u_k^{\mathrm{TSDSC,qp}}$ and $u_k^{\mathrm{TSIDSC,qp}}$ were obtained solving the minimization of quadratic forms (20), (31) using MATLAB function *quadprog*. The performance of these control algorithms is surprisingly essentially inferior with respect to $u_k^{\mathrm{TSDSC,co}}$ and $u_k^{\mathrm{TSIDSC,co}}$. On the other hand, as expected, the control $u_k^{\mathrm{TSIDSC,co}}$ performs better than $u_k^{\mathrm{TSDSC,co}}$.

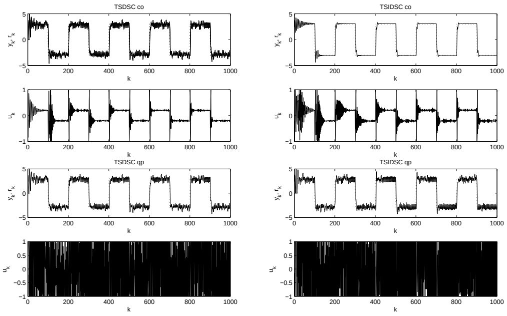


Figure 1: Reference, output and control signals for TSDSC, TSIDSC; $\alpha = 1$; constant parameters.

For the control policy Π^{TSIDSC} the constant learning weight was $\lambda_k = \lambda = 0.98$.

The performance index

$$\bar{J} = \sum_{k=0}^{N-1} (y_{k+1} - r_{k+1})^2$$

was considered for simulations. The plots of $\bar{J}$ versus the constraint α are shown in Fig.2 for $\sigma_w^2 = 0.05$, and $N = 1000$.

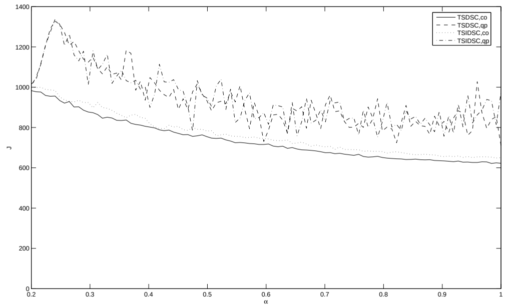


Figure 2: Plots of performance indices for TSDSC, TSIDSC.

In the case of varying parameters (2), $\Phi = I$ and $R_e = 0.05I$ have been taken. Fig.3 shows the performance of the tracking process under the constraint $\alpha = 1$ for both TSDSC and TSIDSC control policies. An examplary run of parameter estimates is shown in Figs.4,5 for control policies TSIDSC, co and TSIDSC, qp, respectively.

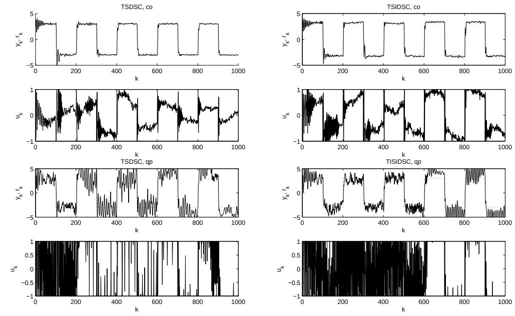


Figure 3: Reference, output and control signals for TSDSC, TSIDSC; $\alpha = 1$; varying parameters.

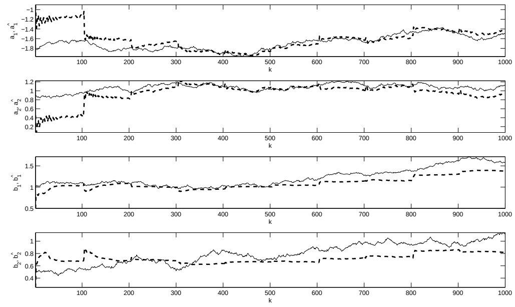


Figure 4: Parameter estimates for TSIDSC, co.

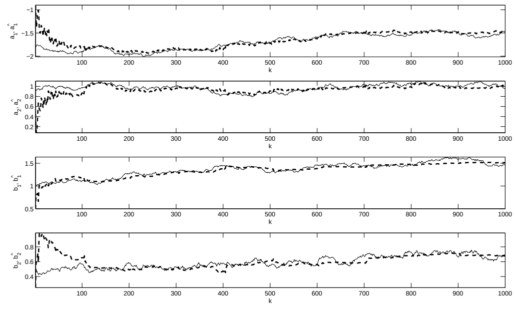


Figure 5: Parameter estimates for TSIDSC, qp.

5 CONCLUSIONS

This paper presents a problem of discrete-time dual control under the amplitude-constrained control signal. A simulation example of second-order system is given and the performance of the presented two control policies is compared by means of the simulated performance index.

A new control policy TSIDSC was proposed as suboptimal dual control approach. The method exhibits good tracking properties for both constant and time-varying unknown system parameters.

It was found that both control policies $u_k^{\text{TSDSC,qp}}$ and $u_k^{\text{TSIDSC,qp}}$ derived via QP optimization do not yield a tracking improvement compared to the cut-off controls $u_k^{\text{TSDSC,co}}$ and $u_k^{\text{TSIDSC,co}}$.

REFERENCES

Åström, H. and Wittenmark, B. (1989). *Adaptive Control*. Addison-Wesley.

Bayard, D. (1991). A forward method for optimal stochastic nonlinear and adaptive control. *IEEE Trans. Automat. Contr.*, 9:1046–1053.

Bayard, D. and Eslami, M. (1985). Implicit dual control for general stochastic systems. *Opt. Contr. Appl.& Methods*, 6:265–279.

D. Li, F. Q. and Fu, P. Optimal nomial dual control for discrete-time linear-quadratic gaussian problems with unknown parameters. *Automatica*, 44:119–127.

Królikowski, A. (2000). Suboptimal lqg discrete-time control with amplitude-constrained input: dual versus non-dual approach. *European J. Control*, 6:68–76.

Królikowski, A. and Horla, D. (2007). Dual controllers for discrete-time stochastic amplitude-constrained systems. In *Proceedings of the 4th Int. Conf. on Informatics in Control, Automation and Robotics*.

Maitelli, A. and Yoneyama, T. (1994). A two-stage dual suboptimal controller for stochastic systems using approximate moments. *Automatica*, 30:1949–1954.

N.M. Filatov, H. U. and Keuchel, U. (1993). Dual pole-placement controller with direct adaptation. *Automatica*, 33(1):113–117.

R. Milito, C.S. Padilla, R. P. and Cadorin, D. (1982). An innovations approach to dual control. *IEEE Trans. Automat. Contr.*, 1:132–137.

Sebald, A. (1979). Toward a computationally efficient optimal solution to the lqg discrete-time dual control problem. *IEEE Trans. Automat. Contr.*, 4:535–540.

Sworder, D. (1966). *Optimal Adaptive Control Systems*. Academic Press, New York.

HIGHER ORDER SLIDING MODE CONTROL FOR CONTINUOUS TIME NONLINEAR SYSTEMS BASED ON OPTIMAL CONTROL

Zhiyu Xi and Tim Hesketh
School of Electrical Engineering & Telecommunications, University of New South Wales, Kensington, NSW, Australia
z3147037@student.unsw.edu.au, t.hesketh@unsw.edu.au

Keywords: Sliding mode, High order, Nonlinear, Optimal control.

Abstract: This paper addresses higher order sliding mode control for continuous nonlinear systems. We propose a new method of reaching control design while the sliding surface and equivalent control can be designed conventionally. The deviations of the sliding variable and its high order derivatives from zero are penalized. This is realized by minimizing the amplitudes of the higher order derivatives of the sliding variable. An illustrative example— a field-controlled DC motor— is given at the end.

1 INTRODUCTION

Variable structure systems (VSS) have been extensively used for control of dynamic industrial processes. The essence of variable structure control (VSC) is to use a high speed switching control scheme to drive the plant's state trajectory onto a specified and user chosen surface in the state space which is commonly called the sliding surface or switching surface, and then to keep the plant's state trajectory moving along this surface (Utkin, 1992), (Utkin, 1977). VSS has attracted attention during the past decades because of its superior capability to eliminate the impact of uncertainties.

Standard sliding mode controllers reveal drawbacks: high frequency vibration of the controlled system, which is also called "chattering", and sensitivity to disturbances during reaching mode. In recent years, a new method, so-called "higher order sliding mode (HOSM)" has been proposed (Levant, 1996), (Levant, 2007), (Glumineaus, 2006) for nonlinear sliding mode design. In higher order sliding mode problems, the switching controller also influences the higher order derivatives of the sliding variable rather than just the first order derivative. Under certain circumstances, for instance, the control u is treated as an additional state variable while its derivative $\dot{u}$ is employed as the actual control (Levant, 1996), (Zinober). The most popular higher order sliding mode controllers are the so called "twisting controller" and "super-twisting controller" which are derived based on bang-bang control theory. A number of papers report the derivation and performance of these controllers (Levant, 1996), (Levant, 2007), (Glumineaus, 2006), (Castellanos, 2004). As discussed by Boiko, Fridman and Castellanos (2004), if the actuator is of second or higher order there is an opportunity for reduction of the amplitude of chattering in the control signal when using twisting as a filter algorithm, compared with first order SM control. In other words, higher order sliding mode control contributes to suppressing the chattering effect although not completely eliminating it. Furthermore, a new concept, "integral sliding mode control (ISMC)" has been developed recently (Shi, 1996). With an integral sliding mode control scheme, the reaching phase is eliminated so that robustness is guaranteed right from the initial time instant.

The aim of this paper is to provide an effective and more convenient way to solve nonlinear higher order sliding mode problems. Nonlinear continuous systems are worked on and second or even higher order sliding mode control concepts are developed. With this method, a sliding mode is reached by forcing the sliding variable and its higher order derivatives to zero in finite time rather than working on nonlinear inequalities based on high order differential equations, which is inevitable in "super-twisting" controller design. The resulting reaching controller does not contain any high frequency switching component which evokes high frequency responses of the system. This idea is borrowed from optimal control laws. The derivation of equivalent control is different from that of normal sliding mode. Meanwhile, the sliding surface design may employ various methods.

In Section III B, we address the problem of the reachability of the sliding surface. To avoid chattering, whatever the initial state of the system is, both the sliding variable and its derivatives have to be driven to zero (not necessarily with the same convergence speed). They should also be kept at zero after the sliding surface is reached. In this paper, the reaching controller is expected to be a continuous nonlinear function with respect to the state variables. The form of the nonlinearity is determined by the solution of a minimization problem which is analogous to that which occurs in optimal control. What is to be minimized is the amplitude of the vector the entries of which are the sliding variable and its derivatives. If an q -th order sliding mode is pursued, the sliding variable and its derivatives up to order $(q-1)$ will be contained in the state vector. This method leads to a very smooth system trajectory. The reachablility of the sliding surface is guaranteed by the existence of a solution to the minimization operation. Furthermore, the minimization algorithm promises good robustness while the precision of high order sliding mode is kept.

At the end of this paper, a field controlled DC motor is considered. The performance of the proposed control scheme is shown applied to this third order system.

2 THE PROBLEM STATEMENT

Consider a continuous nonlinear system of the form

$$
\begin{aligned}
\dot{x}(t) &= f(x(t))+Bu(t), t \geq t_0 && (1)\\
x(t_0) &= x_{0.} && (2)
\end{aligned}
$$

where $x(t) \in R^{n\times 1}$, $u \in R^{1\times 1}$ is the control input, σ is the sliding variable. $B^{n\times 1}, C, D$ are matrices or vectors of proper dimensions and n is known. It is assumed that $f(x(t))$ is Lipschitz continuous and differentiable with respect to $x(t)$ to any order. In this paper, the sliding variable is restrained to be a linear combination of the states, which has the following form:

$$
\sigma(t) = Sx(t) = s_1 x_1(t) + s_2 x_2(t) + ... + s_n x_n(t) \quad (3)
$$

Calculate the first and second order derivative of the sliding variable and we have

$$
\dot{\sigma}(t) = S\dot{x}(t) = Sf(x(t)) + SBu(t) \quad (4)
$$

$$
\begin{aligned}
\ddot{\sigma}(t) = S\ddot{x}(t) = S\frac{\partial f(x(t))}{\partial x(t)} f(x(t))\\
+S\frac{\partial f(x(t))}{\partial x(t)} Bu(t) + SB\dot{u}(t) \quad (5)
\end{aligned}
$$

3 SECOND ORDER SLIDING MODE CONTROL DESIGN

3.1 Sliding Surface Design

For system (1), perform a similarity transformation defined by an orthogonal matrix $P^{n\times n}$:

$$
x_l = Px = [x_{l1} : x_{l2}]^T,\ B_l = PB = \begin{bmatrix} 0_{k\times 1} \\ B_2 \end{bmatrix}, \quad (6)
$$

$$
f_l(x(t)) = f(x_l(t)) = \begin{bmatrix} f_{l1}(x(t)) \\ f_{l2}(x(t)) \end{bmatrix}. \quad (7)
$$

where $x_{l1} \in R^{k\times 1}, x_{l2} \in R^{(n-k)\times 1}, B_2^{(n-k)\times 1}$ and x_{l1} does not have direct dependence on the input. Sliding surface design may be undertaken considering only x_{l1}, treating x_{l2} as an “input” to the partitioned system. In this way, the input may be ignored while determining the sliding surface and this reduces the complexity of the sliding surface design.

The partitioned state equations corresponding to (1) may now be expressed in the following way:

$$
\begin{aligned}
\dot{x_{l1}}(t) &= f_{l1}(x_{l1}(t), x_{l2}(t)) && (8)\\
\dot{x_{l2}}(t) &= f_{l2}(x_{l1}(t), x_{l2}(t)) + B_2 u(t). && (9)
\end{aligned}
$$

Suppose

$$
\begin{aligned}
Sx(t) &= [\begin{matrix} s_1 & s_2 & \cdots & s_n \end{matrix}]x(t)\\
&= wPx(t) = wx_l(t)\\
&= w_{l1}x_{l1}(t) + w_{l2}x_{l2}(t)
\end{aligned}
$$

in which

$$
\begin{aligned}
w_{l1}^{p\times k} &= \begin{bmatrix} w_1 & w_2 & \cdots & w_k \end{bmatrix},\\
w_{l2}^{p\times (n-k)} &= \begin{bmatrix} w_{k+1} & w_{k+2} & \cdots & w_n \end{bmatrix},
\end{aligned}
$$

and $Sx(t)$ is the sliding variable, then the condition for the sliding mode to exist is

$$
w_{l1}x_{l1}(t) + w_{l2}x_{l2}(t) = 0,
$$

which yields

$$
x_{l2}(t) = -w_{l2}^{-1} w_{l1} x_{l1}(t). \quad (10)
$$

When w_{l2} is non-square, w_{l2}^{-1} in (10) should be its pseudo inverse.

Substituting (10) into (8) we have,

$$
\begin{aligned}
x_{l1}(t+1) &= f_{l1}(x_{l1}(t), -w_2^{-1} w_1 x_{l1}(t)) && (11)\\
&= F(x_{l1}(t)) && (12)
\end{aligned}
$$

where $F(\cdot)$ denotes the nonlinear function about $x_{l1}(t)$ after tidying (11) up.

The goal of the next step is to fix the relationship between $x_{l2}(t)$ and $x_{l1}(t)$ to prescribe desirable performance for the nominal sliding mode dynamics. Any standard design algorithm which produces a linear state feedback controller for a nonlinear dynamic system can be used to determine $F(x_{l1}(t))$ and achieve desired performance through selection of sliding mode dynamics (Spurgeon, 1992). It is also assumed here that (12) is stabilizable. The controller gain derived is:

$$x_{l2}(t) = -kx_{l1}(t) \tag{13}$$

which means that

$$\sigma(\mathbf{x}_l(t)) = \begin{bmatrix} k & \vdots & I \end{bmatrix} x_l(t) \tag{14}$$

while I represents the identity matrix with proper dimension.

Note that inversion of the similarity transformation (using P) is needed to recover $x(t)$ from $x_l(t)$. Then $Sx(t) = 0$ is the desired sliding surface.

3.2 Higher Order Sliding Mode Design

3.2.1 Reaching Control Design

As the reaching condition implies, the sliding variable has to converge to zero in finite time. Furthermore, as an q-th order sliding mode is expected, $\dot{\sigma}$, $\ddot{\sigma}$......$\sigma^{(q-1)}$ are also desired to approach zero. Derive a vector containing $\sigma, \dot{\sigma}, \ddot{\sigma}$......$\sigma^{(q-1)}$ and extend (4), (5) to describe this vector

$$\dot{\sigma}(t) = S\dot{x}(t) = Sf(x(t)) + SBu(t) \tag{15}$$

$$\ddot{\sigma}(t) = S\ddot{x}(t) = S\frac{\partial f(x(t))}{\partial x(t)}f(x(t))$$

$$+S\frac{\partial f(x(t))}{\partial x(t)}Bu(t) + SB\dot{u}(t) \tag{16}$$

$$\sigma^{(3)}(t) = Sx^{(3)}(t)$$

$$= \quad S(\frac{\partial^2 f(x(t))}{\partial x^{(2)}(t)} + (\frac{\partial f(x(t))}{\partial x(t)})^2)f(x(t)$$

$$+S(\frac{\partial^2 f(x(t))}{\partial x^{(2)}(t)} + (\frac{\partial f(x(t))}{\partial x(t)})^2)Bu(t)$$

$$+S\frac{\partial f(x(t))}{\partial(t)}B\dot{u}(t) + SBu^{(2)}(t), \tag{17}$$

$$\ldots\ldots$$

which is equivalent to

$$z(t) = G(x(t)) + H(x(t))U(t). \tag{18}$$

where

$$z(t) = \begin{bmatrix} \sigma(t) \\ \dot{\sigma}(t) \\ \ldots \\ \sigma^{(q-1)}(t) \end{bmatrix}, \; U(t) = \begin{bmatrix} u(t) \\ \dot{u}(t) \\ \ldots \\ u^{(q-1)}(t) \end{bmatrix},$$

$$H(x(t)) =$$

$$\begin{bmatrix} 0 & 0 & \ldots & 0 \\ SB & 0 & \ldots & 0 \\ S\frac{\partial f(x(t))}{\partial x(t)}B & SB & \ldots & 0 \\ S(\frac{\partial^2 f(x(t))}{\partial x^{(2)}(t)} + (\frac{\partial f(x(t))}{\partial x(t)})^2)B & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & SB \end{bmatrix},$$

$$G(x(t)) = \begin{bmatrix} Sx(t) \\ Sf(x(t)) \\ S\frac{\partial f(x(t))}{\partial x(t)}f(x(t)) \\ S(\frac{\partial^2 f(x(t))}{\partial x^{(2)}(t)} + (\frac{\partial f(x(t))}{\partial x(t)})^2)f(x(t)) \\ \ldots \end{bmatrix} \tag{19}$$

Here, the following conditions are assumed:

Assumption I: $z(t) \in Z, Z$ contains the origin.

Assumption II: The set Z is reachable in finite time from any initial state and from any point in the generated trajectories.

As the purpose of reaching control design is to find some $u(t)$ which regulates $z(t)$ to zero in finite time, we define a cost function which is

$$J(t) = z^T(t)z(t) + \lambda U^T(t)U(t) \tag{20}$$

with a weighting factor λ. Then $U(t)$ is determined to minimize $J(t)$.

Taking the partial derivative of $J(t)$ with respect to $U(t)$ we have:

$$\frac{\partial J(t)}{\partial U(t)} = \frac{\partial(z^T(t)z(t) + \lambda U^T(t)U(t))}{\partial U(t)}. \tag{21}$$

Let $\frac{\partial J(t)}{\partial U(t)} = 0$ and derive:

$$U(t) = M(x(t))$$

$$= -(H(x(t))^T H(x(t)) + \lambda I)^{-1} H(x(t))^T G(x(t)) \tag{22}$$

(I here again represents the identity matrix if certain dimension.)

It should be noticed that the derivation of (22) reduces to a Tikhonov regularization problem therefore the detail is omitted here.

REMARK: Here, we assume the minimization over an infinite horizon results in a control $U^*(t)$. This control input will be implemented only until the next measurement becomes available. Then the up to date system information will be taken into account and a new value of $U^*(t)$ is computed. Introduce

$$\begin{aligned} J(t+h) &= z^T(t)z(t)+\lambda U^{*T}(t)U^*(t) && (23)\\ &\le z^T(t)z(t)+\lambda U^T(t)U(t)=J(t)(24) \end{aligned}$$

where $J(t)$ stands for the cost observed at time t and h is a sufficiently small positive number. The final cost $J(\infty)$ is a finite non-negative number as $J(t)$ is non-increasing. In other words, $J(t)$ decreases due to the effect of $U^*(t)$ until reaches zero. Then the next value (final value) of $U^*(t)$ is zero which indicates that the reaching mode is complete. Meanwhile, the final value of $z^T(t)z(t)$ is zero. By choosing λ to be a small positive weighting factor, non-zero $z^T(t)z(t)$ will be relatively heavily punished and so $z(t)$ converges to zero more quickly.

The reaching control law $u_r(t)$ can be obtained from the equation which forms the first row of (22) (Wertz, 1990)

$$u_r(t)=\begin{bmatrix} 1 & 0 & \dots & \dots & 0 \end{bmatrix} M(x(t))=M_1(x(t)) \quad (25)$$

where $M_1(x(t))$ stands for the first element of vector $M(x(t))$.

3.2.2 Robustness Issue

By substituting (22) into (18) we have

$$z(t)=G(x(t))+H(x(t))M(x(t)). \quad (26)$$

Now, assume that due to modelling errors, the real system is

$$\dot{x}(t)=f_{real}(x(t))+B_{real}u(t), t\ge t_0 \quad (27)$$

which leads to

$$z(t)=G_{real}(x(t))+H_{real}(x(t))U(t). \quad (28)$$

The robustness of the reaching mode relies on

- *Assumption I and II* for $z(t)$ in (28)
- The satisfaction of (29)

$$J(G_{real},H_{real},U^*(t),t)\le J(G,H,t). \quad (29)$$

3.2.3 Equivalent Control Design

After the sliding mode is reached, the system dynamic is dominated by the equivalent controller. To ensure $q-th$ order sliding, the equivalent control has to maintain $\sigma(t),\dot{\sigma}(t)...\sigma^{(q-1)}(t)$ at zero. By extending (1), (3) we have

$$\sigma^{(q-1)}(t)=P(f(x))+Q(u(t)). \quad (30)$$

where $P(\cdot)$ and $Q(\cdot)$ are both nonlinear functions.

The equivalent control $u_{eq}(t)$ should be derived according to the following

$$\sigma^{(q-1)}(t)=P(f(x))+Q(u_{eq}(t))=0. \quad (31)$$

As introduced in (Matthews, 1988), the complete sliding mode controller is

$$u(t)=u_{eq}(t)+u_r(t) \quad (32)$$

where $u_r(t)$ is from (25).

4 EXAMPLE AND SIMULATION RESULTS

4.1 Field controlled DC Motor and controller Design

Consider the example of a field-controlled DC motor. DC motors are widely used by almost all industries and can be highly nonlinear in field controlled configurations. The mathematical model of a DC motor can be expressed in the following way.

Figure 1: Structure of a DC motor.

$$\begin{aligned} \dot{x_1}(t) &= -ax_1(t)+u(t) && (33)\\ \dot{x_2}(t) &= -bx_2(t)+\rho-cx_1(t)x_3(t) && (34)\\ \dot{x_3}(t) &= \theta x_1(t)x_2(t) && (35) \end{aligned}$$

$$y(t)=x_3(t). \quad (36)$$

The physical meanings of the variables in the above equations are:

$x_1(t)$	Field current
$x_2(t)$	Armature current
$x_3(t)$	Angular velocity ,
$u(t)$	Field voltage
ρ	Armature voltage,

with a, b, c, θ, ρ positive constants.

The equilibria of the system are

$$x_1 = 0, x_2 = \frac{\rho}{b} \text{ and } x_3 = \omega_0,$$

where ω_0 is a desired setpoint for the angular velocity.

In this paper, we choose

$$a = b = c = \theta = \rho = 1$$

for simplicity (**?**).

The partitioned system matrices are

$$\begin{aligned} f_{l1}(x_{l1}, x_{l1}) &= -x_1(t) && (37)\\ f_{l2}(\dot{x_{l1}}, x_{l1}) &= \begin{bmatrix} -x_2(t) + 1 - x_1(t)x_3(t) \\ x_1(t)x_2(t) \end{bmatrix} && (38)\\ B_l &= \begin{bmatrix} B_1 \\ 0_{2\times1} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} && (39) \end{aligned}$$

It is seen that (33)-(35) are already in the same form as (6)-(7). Hence the transformation matrix P is identity.

Suppose

$$Sx(t) = [\ s_1 \quad s_2 \quad s_3\ ]x(t) = w_{l1}x_{l1}(t) + w_{l2}x_{l2}(t),$$

The values of w_1 and w_2 must be chosen to ensure the following system has satisfactory closed loop behavior:

$$x_{l2}(t) = Kx_{l1}(t) = -w_{l2}^{-1}w_{l1}x_{l1}(t)$$

$$\begin{aligned} \dot{x_{l1}}(t) &= f_{l1}(x_{l1}(t), -w_{l2}^{-1}w_{l1}x_{l1}(t)) && (40)\\ &= Fx_{l1}(t). && (41) \end{aligned}$$

One of the proper selections of w_1 and w_2 leads to:

$$K = [\ 0 \quad -1\ ]$$

which produces a sliding variable:

$$\sigma(t) = Sx(t) = x_1(t) + x_3(t) - x_{3desired}$$

In this case, $x_3(t) - x_{3desired}$ is treated as the state of the system rather than $x_3(t)$ in certain design steps because the final value of x_3 is not expected to be zero but a desired value. This desired value should be involved in the sliding surface design. Similarly, $x_{2desired}$ should be considered at some stage as well. (In this case, $x_{2desired} = 0.95$ and $x_{3desired} = 2.05$.) Accordingly, we have

$$u_{eq}(t) = \frac{2x_1^2(t)x_3(t) + 2x_1(t)x_3(t) - x_1^3(t)x_2(t) - 4x_1(t)}{3x_2(t) + x_1(t)x_3(t) + 2x_3(t) - 3}$$

which is derived by letting

$$\sigma^{(3)}(t) = Sx^{(3)}(t) = 0.$$

Now we proceed to design the reaching control. In this case, take the derivatives of $\sigma(t)$ up to order 3 into account in the cost function definition. Then $G(x(t))$ and $H(x(t))$ can be calculated from (19). As the result of the minimization, the reaching control will be expressed in the form of a nonlinear function of state variables:

$$u_r(t) = W(x(t)).$$

$W(x(t)$ is derived from (22) and (25). Computing $W(x(t))$ is reduced to a numerical calculation without necessity of pursuing the algebraic description of $W(x(t))$. Finally the complete control law $u(t)$ is derived using (32) with the equivalent control derived according to (31).

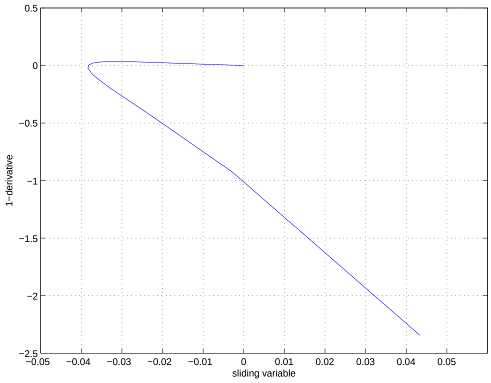


Figure 2: Sliding variable and its derivative.

4.2 Simulation Results

The integration step size is chosen to be $1ms$. In all the figures below, the unit of time axis is in second. The process depicting the sliding variable and its derivative as they approach zero is shown in Fig. 2 The trajectory travels smoothly on the plane until it reaches the origin without overshooting. From Fig. 3

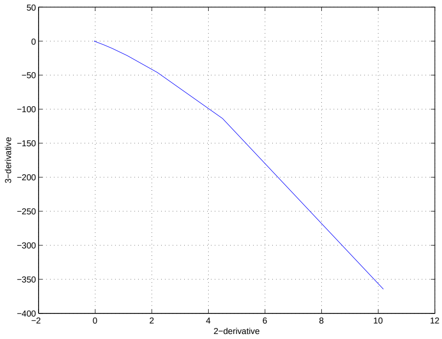


Figure 3: Higher order derivatives of the sliding variable.

we can see that the second and third order derivatives of the sliding variable also behave as a smooth curve which ends up at the origin. Figure 4 shows the convergence performance of the state variables. It is shown that $x_1(t)$ converges to zero while $x_2(t)$ and $x_3(t)$ each approach their desired value. The trajectories are smooth and there is no overshoot or oscillation. The whole process settles quickly within 0.5 seconds.

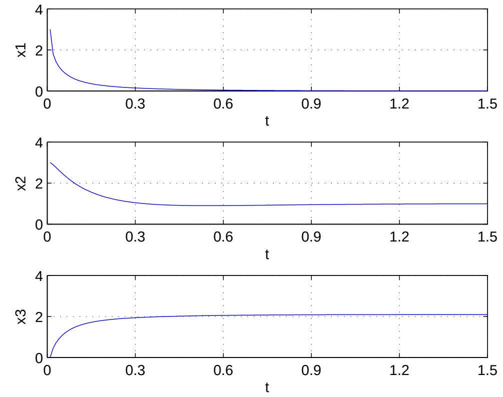


Figure 4: Convergence of the states.

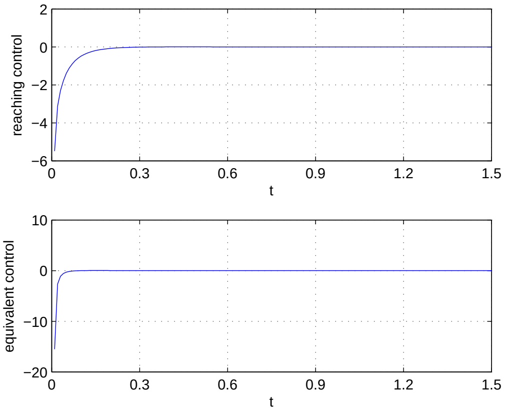


Figure 5: Control signal u.

The variation of the control signal u during the period is plotted in Figure 5.

As shown above, a good performance is achieved. A higher order sliding behavior is shown.

5 CONCLUSIONS

In this paper, a new method of designing a higher order sliding mode controller for a continuous nonlinear dynamic system is reported. Retaining the advantages of higher order sliding mode control, i.e. chattering reduction, the complexity of nonlinear design is greatly reduced with this method especially in reaching control design. A field-controlled DC motor is given as an illustrative example to show the effectiveness of this method.

REFERENCES

Vadim I. Utkin (1992). *Sliding Modes in Control and Optimization*. Springer-Verlag New York, Inc.

Vadim I. Utkin (1977). *Variable Structure Systems with Sliding Modes*. IEEE Transactions on Automatic Control, Volume AC-22, No. 2, April.

S. V. Emeryanov, S. K. Korovin, and A. Levant (1996). *Higher-order Sliding Modes in Control Systems*. Computational Mathematics and Modeling, Vol. 7, No. 3.

A. Levant (2007). *Principles of 2-sliding mode design*. Automatica Vol. 43, pp. 576 – 586.

S. Laghrouche, F. Plestan, and A. Glumineau (2005). *Multivariable practical higher order sliding mode control*. Proceedings of the 44th IEEE Conference on Decision and Control, and the European Control Conference.

A.J. Koshkouei, K.J. Burnham and A.S.I. Zinober. *Dynamic sliding mode control design*. IEE Proceedings online no. 20055133.

I. Boiko, L. Fridman, and I. M. Castellanos (2004). *Analysis of second-order sliding-mode algorithms in the frequency domain*. IEEE Transaction on Automatic Control, Vol. 49, No. 6, pp. 946–950, Jun.

Vadim Utkin and Jingxin Shi(1996). *Integral Sliding Mode in Systems Operating under Uncertainty Conditions*. Proceedings of the 35th Conference on Decision and Control, Kobe, Japan, December.

S. K. Spurgeon (2004). *Temperature Control of Industrial Process using a Variable Structure Design Philosophy*. Transactions of the Institute of Measurement and Control

R. R. Bitmead, M. Gevers, and V. Wertz (1990). *Adaptive Optimal Control: The Thinking Man's GPC*. Englewood Cliffs, NJ: Prentice-Hall.

Raymond, A. DeCarlo, Stanislaw H. Zak, Gregory P. Matthews(1988). *Variable Structure Control of Nonlinear Multivariable Systems: A Tutorial*. Proceedings of The IEEE, Vol. 76, No. 3, March.

Christopher Edwards and Sarah K. Spurgeon (1998). *Sliding Mode Control, Theory And Applications*. CRC Press, Taylor & Francis Group.

AN ANALYTICAL AND NUMERICAL STUDY OF PRESSURE TRANSIENTS IN PNEUMATIC DUCTS WITH FINITE VOLUME ENDS

N. I. Giannoccaro, A. Messina and G. Rollo
Dipartimento di Ingegneria dell'Innovazione, Università del Salento, Via per Monteroni, Lecce, Italia
ivan.giannoccaro@unile.it, arcangelo.messina@unisalento.it, giosue.rollo@unile.it

Keywords: Industrial automation, Pneumatic transmission line, Finite volumes, Ducts, Fluid mechanics.

Abstract: In this paper, the response of a pneumatic transmission line is analysed through two different approaches. Both the approaches, based on the same physical model, are able to simulate the dynamics of a pneumatic line, with finite volume ends. The first approach analytically provides the transients through an equation in a quasi-closed form; the second approach is based on a numerical procedure yielding the inversion of the Laplace transform by the application of a trapezoidal rule. The analysis of the mutual performances of the two approaches, in the frame of pneumatic systems normally operating in industrial automation, can be useful in terms of control of the response and could assist in the design of pneumatic systems.

1 INTRODUCTION

Pneumatic actuators are often employed in industrial automation for reasons related to their good power/ weight ratio, easy maintenance and assembly operations, clean operating conditions and low cost.

This set of advantages, however, is negatively balanced by the difficulties met during the design. Indeed, the presence of air, along with its natural compressibility, introduces further complexities to those already existing: friction forces, losses and time delays in cylinder and transmission lines (Messina, 2005), (Carducci, 2006). For these reasons, fast transients involved in wave propagations in pneumatic transmission lines deserve to be taken into account in the design of the system (Rollo, 2007).

The pneumatic transmission line, analysed in this work, consists of a tube, of a certain length, connecting two finite capacities. As far as the gas-dynamic inside the duct is concerned, in literature there are several papers dealing with such systems but only describing lines with one of the two capacities being finite. In this work, the mathematical description of a pneumatic line, with finite volume terminations, is presented. This setting complicates the mathematics of the phenomena but it is interesting because of the industrial practice (Messina, 2005), (Rollo, 2007) where finite capacities are very common. The transient in the line is described through two partial differential equations, whose solution is obtained in correspondence to suitable initial and boundary conditions. The mathematical model (Rollo, 2007) gives the pressure response for a double volume terminated pneumatic line and includes as a particular case a previous model presented by Schuder and Binder (Schuder, 1959).

The model is obtained assuming small pressure and temperature changes, such that the following assumptions are valuable (Schuder, 1959): (i) incompressible flow and (ii) laminar flow; the accuracy of the response, in correspondence of different operative conditions, has been discussed elsewhere (Rollo et al., 2007).

The assumptions of the model mainly concern the flow conditions which allow an approach based on the Laplace transform (Rollo, 2007), (Schuder, 1959). This type of model could be considered attractive in the frame of industrial automation, but a possible difficulty arises in the inverse transformation especially when an analytical description of the transient is attempted (Rollo, 2007). In this respect, a numerical method (Crump, 1976), (Duffy, 1993), that readily determines the Laplace transform inversion, could be considered attractive in order to achieve the pressure transient.

These two approaches, the analytical one (Rollo, 2007) and the numerical one (Duffy, 1993), that have different complexities, are taken into account herein. The first analytically provides the description of the pressure transients through an equation in a quasi-closed form. The second is numerically able to yield the inversion of the Laplace transform in a direct way, through a trapezoidal rule (Crump, 1976), (Duffy, 1993). This latter, under certain conditions, requires no manipulation on Laplace transform.

The two approaches, with the analysis of the mutual performances, can suggest, in the frame of pneumatic systems normally operating in industrial automation, the strategies in terms of design and control of the response. In this respect, interesting conclusions can be extracted.

2 SYSTEM ANALYSED

For a self comprehension of the present work, a brief description of the real system analysed is also presented. The relevant physical model which is referred to in the present work is illustrated in Fig. 1.

The system under investigation consists of two chambers having volumes Q_1, Q_2. The chambers are connected through a cylindrical tube (also termed as pneumatic transmission line) whose transversal section is constant in the range of commercial tolerances. The x-longitudinal coordinate is settled from the upstream (chamber 1: Q_1) to the downstream chamber (chamber 2: Q_2).

The upstream chamber consists of a five litre tank arranged with four holes in order to allow the external connections. In particular, chamber 1 is filled up through a tap air supply until an established static pressure, measured by the absolute pressure gage, is reached. An airtight adapter is screwed onto chamber 1. The adapter is made airtight through an internal membrane made of commercial sticky tape.

The test and simulated condition consists of suddenly breaking the membrane in order to allow a wave pressure travel from chamber 1 to chamber 2 and vice versa; the sudden rupture of the membrane is caused by a puncturing actuator placed at the symmetrical end with respect to the membrane; the puncturing actuator is quasi-statically activated by manually pushing its rod through orifice A.

When a step pressure signal propagates through the duct an on/off valve can be considered simulated (Rollo, 2007). Based on these motivations the approaches (analytical and numerical) have been tested with respect to the mentioned step-type signal.

The downstream volume consists of the ram chamber of a commercial double acting pneumatic actuator. The established volume Q_2 can be in practice settled by grounding the rod of the actuator at a fixed position.

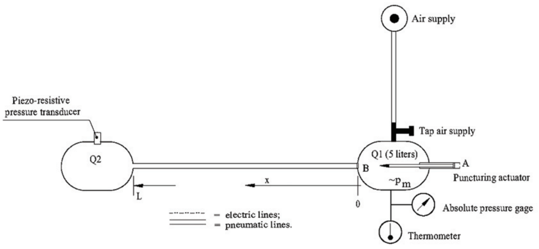


Figure 1: Scheme and nomenclature of the system.

3 SCHUDER AND BINDER EXTENDED (SBE) MODEL: ANALYTICAL APPROACH

In the SBE model (Rollo, 2007), (Schuder, 1959), the equation describing the pressure transient in the duct, obtained from an analytical solution of two partial differential equations (one-dimensional mass and momentum conservation law (Schuder, 1959)), is obtained using the Laplace transform.

This kind of procedure yields, in the Laplace domain and in correspondence of an established section of the line (x=L), the following response:

$$P(L;s)=\left(\frac{p_m-p_0}{s}\right)\cdot kQ_1\cdot \frac{1}{\left[\begin{array}{l}(kQ_1+Q_2)\cosh\left(L\sqrt{\beta}\right)+\\ \left(\frac{a}{\sqrt{\beta}}+\frac{\sqrt{\beta}}{a}kQ_1Q_2\right)\sinh\left(L\sqrt{\beta}\right)\end{array}\right]}+\frac{p_0}{s} \qquad (1)$$

where a is the cross sectional area of the duct, k the ratio of specific heats at pressure and volume constant (c_p/c_v), L the length of the duct, p_m and p_0 the initial pressure in the sending volume and connecting duct respectively, Q_1 and Q_2 the sending and receiving volume respectively and β is the following parameter:

$$\beta=\frac{s(R+\rho s)}{\rho c^2} \qquad (2)$$

depending on R (frictional resistance in duct in the presence of laminar flow), s the Laplace variable, ρ the density (constant) and c the sound speed (constant).

The inverse transform of (1) is not a straightforward task; in this respect, following Schuder and Binder, Jaeger's result is taken into account (Schuder, 1959): evaluating the coefficients of an exponential series, it is possible to obtain the following analytical solution showing the pressure in the time domain in the position close to chamber 2 (x=L):

$$P(L;t)_A = \frac{p_m kQ_1 + p_0(Q_2 + aL)}{kQ_1 + Q_2 + aL} - 2(p_m - p_0)e^{-\frac{Rt}{2\rho}} \cdot$$

$$\cdot \sum_{n=1}^{\infty} \frac{\left[\cos\left(\frac{\vartheta_n t}{2}\right) + \frac{R}{\rho\vartheta_n}\sin\left(\frac{\vartheta_n t}{2}\right)\right]}{\alpha_n \left[\begin{array}{l}\left(1 + \frac{Q_2}{kQ_1} + \frac{Q_2}{aL} + \frac{aL}{\alpha_n{}^2 kQ_1}\right)\sin\alpha_n + \\ \left(\frac{Q_2\alpha_n}{aL} - \frac{aL}{\alpha_n kQ_1}\right)\cos\alpha_n\end{array}\right]} \tag{3}$$

where t is the time; α_n and θ_n are functions of the geometry of the line and initial flow conditions; in particular, the poles in Equation (1) are function of α_n too (Rollo, 2007), (Schuder, 1959). In fact, by substituing in Equation (1)

$$L\sqrt{\beta} = L\left(\frac{s(R+\rho s)}{\rho c^2}\right)^{\frac{1}{2}} = i\alpha \tag{4}$$

and equating to zero the common denominator of Equation (1), an implicit equation in α is obtained; once it is solved (numerically, with the Newton-Raphson method), after some simple mathematical manipulations, it is possible to obtain, along with $s_0=0$ (Rollo, 2007), (Schuder, 1959):

$$s_n = -\frac{R}{2\rho} \pm \frac{i}{2}\sqrt{\left(\frac{2\alpha_n c}{L}\right)^2 - \left(\frac{R}{\rho}\right)^2} \tag{5}$$

in which the real part of the poles is negative or zero.

The quasi-closed form solution (3) depends on the number of the terms in the series; the only non-serious drawback is the necessity of resorting to a numerical method in order to assess α_n. Equation (3) yields p(L;t) for a double volume terminated pneumatic line and it is an extension of a previous model presented by Schuder and Binder (Schuder, 1959). The analytical approach to the SBE model is one of the most complete treatments among those presented about transients in pneumatic lines within relevant literature, in which the influence of pressure waves propagating in ducts is, sometimes, neglected or poorly described. Furthermore, even if it is obtained using the following assumptions: (i) incompressible flow and (ii) laminar flow, it can be adopted, without significantly reducing the accuracy of the response, in correspondence to certain operative conditions (Rollo et al., 2007), where the relevant investigations showed its ability to describe pressure transients including reflecting waves.

4 NUMERICAL APPROACH

The Jaeger's results, in the SBE model, are related to the fact that the solution (3) depends i) on the numerical evaluation of α_n and ii) on a certain amount of labor for the mathematical procedure leading to Equation (3) in the inverse Laplace transform (Rollo, 2007), (Schuder, 1959). A kind of resolution allowing to obtain the inverse transform of Equation (1) in a direct way, could be considered attractive in the frame of systems normally operating in industrial automation. In this respect the authors suggest, in this work, to solve the problem of readily determining the inverse Laplace transform using a numerical approach. Within relevant literature, a large number of different methods for numerically inverting the Laplace transform have been introduced and tested: one of these uses a Fourier series approximation (Crump, 1976), (Duffy, 1993). In fact, in (Duffy, 1993) the following straightforward application of a trapezoidal rule in order to provide the numerical inversion of Laplace transform (here referred to Eq. (1)) is proposed:

$$P(L;t) = \frac{e^{\mu t}}{2t}\left\{\begin{array}{l}\frac{1}{2}\mathrm{Re}[P(L;\mu)] + \\ \sum_{z=1}^{\infty}(-1)^z\left[\begin{array}{l}\mathrm{Re}\left[P\left(L;\mu + \frac{z\pi i}{t}\right)\right] + \\ \mathrm{Im}\left[P\left(L;\mu + \frac{(2z-1)\pi i}{2t}\right)\right]\end{array}\right]\end{array}\right\} \tag{6}$$

where t is the time, i the imaginary unit, Im and Re indicate the imaginary and real part of the quantity in the brackets respectively and μ must be greater than the real part of any singularity (poles) in P(L;s) (Duffy, 1993).

The accuracy and efficiency of such a numerical approach depend on a suitable choice of some parameters. In particular, μ can be evaluated through certain considerations about the discretization error (related to the step size π/t for the trapezoidal rule) deriving from the application of Eq. (6). Following the approach proposed in (Crump, 1976), introducing the hypothesis that the function of interest is bounded by:

$$|P(L;t)| \le Me^{\lambda t} \tag{7}$$

(with M and λ real numbers) it is possible to choose μ through the following relation:

$$\mu = \lambda - \frac{\ln(Err)}{2t} \tag{8}$$

where Err is the error parameter within the numerical accuracy desired and the parameter λ can be chosen slightly larger than the maximum of the real part of all the poles (Crump, 1976). Once μ is known, the series (6) can be summed until it converges to the desired number of significant figures (Crump, 1976). Usually the series in (6) can converge slowly (this has been observed in some tests not reported here); furthermore, (Crump, 1976) the use of a sequence accelerator in conjunction with the numerical inversion is recommended, also in order to obtain a reduction of the truncation error (indeed the series in Eq. (6) is not summed to infinity). In this case, following the considerations in (Crump, 1976), (Duffy, 1993), here Wynn's epsilon-algorithm is adopted. More specifically, to accelerate the convergence of the sequence of partial sums in (6) using the epsilon-algorithm, it is possible to calculate them as in (9):

$$\begin{aligned} S_0 &= \frac{1}{2}\operatorname{Re}[P(L;\mu)], \\ S_z &= S_{z-1} + (-1)^z \left[\begin{array}{l} \operatorname{Re}\left[P\left(L;\mu + \dfrac{z\pi i}{t}\right)\right] + \\ \operatorname{Im}\left[P\left(L;\mu + \dfrac{(2z-1)\pi i}{2t}\right)\right] \end{array} \right], \\ z &= 1,...,2N+1. \end{aligned} \tag{9}$$

It is possible, then, to define $\varepsilon_{-1}^{(m)}=0$, $\varepsilon_0^{(m)}=S_m$, m=0, 1,..,2N and then put:

$$\begin{aligned} &\varepsilon_{p+1}^{(m)} = \varepsilon_{p-1}^{(m+1)} + \left[\varepsilon_p^{(m+1)} - \varepsilon_p^{(m)}\right]^{-1}, \\ &p = 0,...,2N \end{aligned} \tag{10}$$

In this way, the sequence $\varepsilon_0^{(0)}$, $\varepsilon_2^{(0)}$, $\varepsilon_4^{(0)}$,.., $\varepsilon_{2N}^{(0)}$ gives better successive approximation to the sum of the series (Crump, 1976). So, Equation (6), including the sequence accelerator, becomes Equation (11):

$$P(L;t)_{NUM} = \frac{e^{\mu t}}{2t} \cdot \varepsilon_{2N}^{(0)} \tag{11}$$

5 ANALYTICAL VS NUMERICAL APPROACH: CONDITIONS, RESULTS AND DISCUSSIONS

The analytical approach presented in Section 3 (Rollo, 2007) was used for an interesting comparison, with respect to a more refined model NLC (for Non Laminar Compressible flow) (Rollo, 2007), (Rollo et al., 2007). This latter model, whose behaviour was validated through experimental investigations (Rollo, 2007) on the physical model of Fig. 1, takes into account i) flow not necessarily laminar and ii) compressible flow. Through a suitable error parameter, the discrepancy on the response of a pneumatic line with established geometrical characteristics (L=2.53m, D=3 or 6 mm, Q_1= 5dm^3), for polyurethane ducts with an assumed internal roughness of 3μm (Rollo, 2007) was estimated for various flow conditions. In both models, the relevant solution can be obtained and displayed, in space (at a fixed time) and in time (at a fixed position); however, the major relevance of the performance in industrial applications (Messina, 2005), (Rollo, 2007) is related to the behaviour of the pressure in the ram chamber of the actuator (x=L). Therefore, only p(L;t) was discussed, in correspondence of the various receiving volumes Q_2 (obtained fixing the stroke of the pneumatic actuator employed as downstream volume, in correspondence of different positions).

This comparison highlighted that, for a fixed geometrical configuration, the error between the SBE and NLC models is as small as the pressure ratio p_m/p_o is close to 1 (Rollo, 2007); in this case the match of SBE response with the NLC curve is very satisfactory. This comparison was made after appropriate convergence tests, that suggested to use n=0,…..,30 in the Equation (3) for the SBE model.

In this work, the comparison between the analytical (3) and numerical (11) approach is discussed, in correspondence of some settings yielding a satisfactory agreement of the SBE with the NLC model. In particular, a line with L=2.53m, D=3 mm and Q_1= 5dm^3 will be considered. The case of interest concerns the transient caused by an upstream initial pressure in Q_1 of 1.1 times higher than initial atmospheric pressure. This setting is completed arranging the actuator employed as downstream volume to establish, in a first case, a volume Q_2 = 1.5 cm^3 (capacity corresponding to the dead space of the ram chamber) and, in a second case, Q_2 = 26.5 cm^3 (Rollo, 2007).

A suitable error parameter is introduced (12) for the detection of discrepancy of the two approaches (3) and (11):

$$\mathrm{ERROR} = \left| \frac{p(L;t)_A - p(L;t)_{NUM}}{p(L;t)_A} \right| \tag{12}$$

in correspondence to various values of the time (10 ms≤t≤150 ms). In applying (11), it is assumed Err=10^{-6} and, for the considerations about Eqs. (5) and (8), it can be assumed λ=0. In this way, the μ value is known in all the time instants considered. The comparison will be made firstly assuming, as far as the numerical approach (11) is concerned, N=25. In this respect, for the first case (Q_2=1.5 cm^3) the following Table 1 was produced:

Table 1: Simulated Pressure with L=2.53 m, D=3 mm, p_m/p_0=1.1, Q_2=1.5 cm^3 (n=30, N=25).

Time Instants (ms)	Analytical Pressure (bar)	Numerical Pressure (bar)	ERROR
10	1.092289	1.092282	6.8E-06
20	1.116303	1.116265	3.4E-05
30	1.101538	1.101535	2.1E-06
40	1.097992	1.097992	3.5E-08
50	1.099557	1.099561	3.7E-06
60	1.099927	1.099927	1.5E-07
70	1.099681	1.099659	1.9E-05
80	1.099693	1.099693	1.7E-07
90	1.099725	1.099726	2.8E-07
100	1.099725	1.099726	1.1E-07
110	1.099723	1.099723	1.1E-07
120	1.099724	1.099724	2.9E-09
130	1.099724	1.099724	1.5E-08
140	1.099724	1.099724	4.5E-09
150	1.099724	1.099724	4.4E-10
Pressure computational time (ms)	~ 5	~ 290	

In Table 1 the first column shows the time instants taken into account, the second and third columns show the pressure values obtained with the two approaches, the analytical one and the numerical one respectively, the fourth column shows the ERROR values. Table 1 shows an agreement of the analytical and numerical results (second and third column) that can be considered very satisfactory; furthermore, it is possible to notice that the computational times (evaluated on a 2.8 GHz Pentium IV using a Matlab routine) show that the analytical approach needs about 5 ms to provide the pressure values in the second column, whilst the numerical computational time is about 290 ms.

As far as the second case (Q_2=26.5 cm^3) is concerned, the following Table 2 was produced:

Table 2: Simulated Pressure with L=2.53 m, D=3 mm, p_m/p_0=1.1, Q_2=26.5 cm^3 (n=30, N=25).

Time Instants (ms)	Analytical Pressure (bar)	Numerical Pressure (bar)	ERROR
10	1.018830	1.018831	3.6E-07
20	1.053451	1.053449	2.1E-06
30	1.076250	1.076250	1.6E-08
40	1.087726	1.087726	1.9E-08
50	1.093537	1.093539	1.6E-06
60	1.096557	1.096557	9.1E-09
70	1.098006	1.098006	3.3E-08
80	1.098703	1.098703	2.0E-07
90	1.099047	1.099047	5.3E-09
100	1.099214	1.099214	1.6E-07
110	1.099294	1.099293	1.1E-06
120	1.099333	1.099333	5.3E-12
130	1.099352	1.099352	5.9E-09
140	1.099361	1.099361	1.4E-09
150	1.099366	1.099366	1.9E-10
Pressure computational time (ms)	~ 5	~ 290	

Table 2 is also able to show a satisfactory agreement of the results of the two approaches (3) and (11) and, as in Table 1, it is possible to notice that the numerical computational time is about 290 ms, whilst the analytical approach needs about 5 ms to provide the pressure values in all the time instants considered.

Table 1 and Table 2 can suggest that the numerical approach (11) is not always advisable in engineering applications in which the performance is required in terms of design and fast control of the response. A possible way to highlight this drawback can be based on the study of the discrepancies with the analytical approach, decreasing N in such a way as to reduce the computational time of the numerical approach (like in Table 3). In Table 3, the first column shows the operative conditions taken into account, the second column the mean of ERROR values and the third the numerical computational times. As can be seen, the computational times decrease if N decreases, but this behaviour still seems far from the more attractive computational times of the analytical approach. Furthermore decreasing the value of N, the mean ERROR value slightly increases.

Table 3: ERROR for a pneumatic line with L=2.53 m, D=3 mm, p_m/p_0=1.1, Q_2=1.5 cm^3 and Q_2=26.5 cm^3, n=30.

	Mean of ERROR values	Numerical computational times (ms)
N=25 Q_2=1.5 cm^3	4.4E-6	290
N=25 Q_2=26.5 cm^3	3.8E-7	290
N=20 Q_2=1.5 cm^3	2.4E-5	204
N=20 Q_2=26.5 cm^3	7.2E-7	204
N=10 Q_2=1.5 cm^3	6.4E-5	77
N=10 Q_2=26.5 cm^3	7.6E-6	77

An estimation of the behaviour of both approaches concerning the transient in the aforementioned pneumatic lines, can be obtained through the following Fig. 2 and Fig. 3, useful also in order to estimate the pressure transients of Eq. (3) and Eq. (11) in terms of design of the line.

Finally, also Figure 4 has been produced. This latter figure has been introduced with the motivation of showing the satisfactory agreement of both the proposed approaches, also in correspondence of an intermediate receiving volume (Q_2=16.5 cm^3).

6 CONCLUSIONS

In this paper, two approaches of different complexities, concerning the dynamics of a pneumatic transmission line with finite volume ends, have been analysed: one analytical and another one numerical. The first one provides the description of the pressure transients through an equation in a quasi-closed form and gives the pressure response for a double volume terminated pneumatic line.

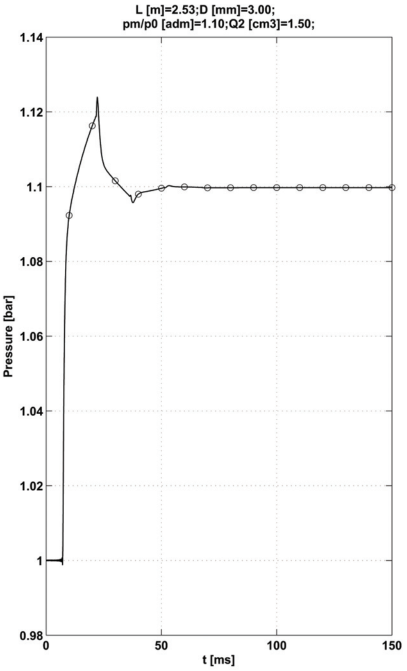


Figure 2: Simulated pressure at the receiving volume through analytical (__) and numerical (°) approach with D=3mm, L=2.53m, p_m/p_0=1.1 and Q_2=1.5cm^3 (n=30, N=25).

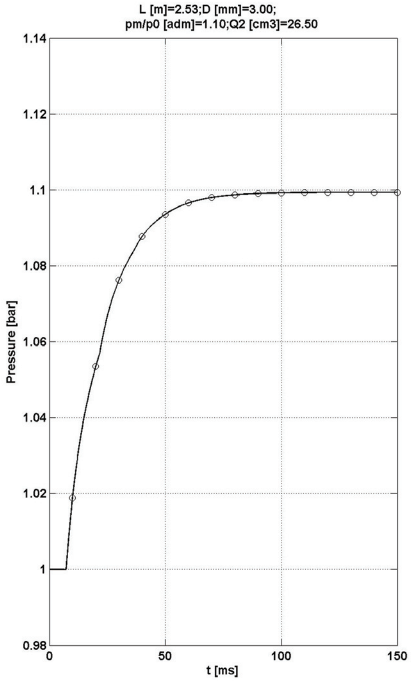


Figure 3: Simulated pressure at the receiving volume through analytical (__) and numerical (°) approach with D=3mm, L=2.53m, p_m/p_0=1.1 and Q_2=26.5cm^3 (n=30, N=25).

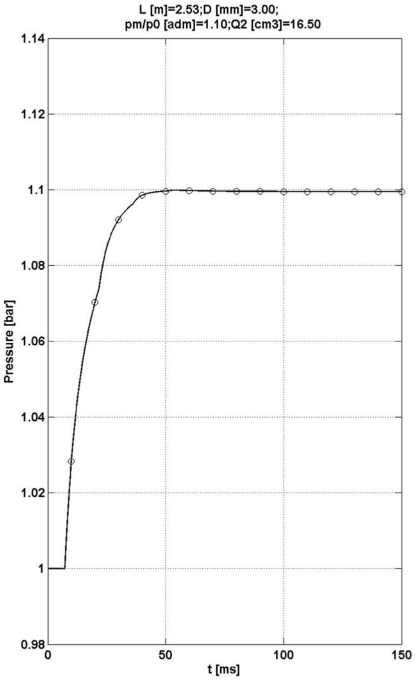


Figure 4: Simulated pressure at the receiving volume through analytical (__) and numerical (°) approach with D=3mm, L=2.53m, p_m/p_0=1.1 and Q_2=16.5cm^3 (n=30, N=25).

The second, numerically, is able to yield the inversion of the Laplace transform through a trapezoidal rule (in conjunction with a sequence accelerator). Using certain geometrical configurations and flow conditions, for which it was shown that the SBE model can be used without significantly reducing the accuracy of the response, the two kinds of resolution can be compared and analysed. The introduction of a suitable error parameter, able to provide the discrepancies of the two approaches, allows interesting discussions.

The trapezoidal rule, in its numerical simplicity, could avoid the long mathematical procedures yielding the inversion of the Laplace transform. The advantage related to the application of this direct rule seems, however, negatively balanced by the extra computational efforts required to achieve a satisfactory convergence (the series approximation can converge slowly and, usually the use of a sequence accelerator in conjunction with the numerical inversion is highly recommended). For these reasons, the numerical approach could not always be advisable in engineering applications in which performances are required in terms of design and control of the response. This, indeed, has been showed by the comparison with an analytical solution in a quasi-closed form, in terms of computational times. The possible drawback of this latter approach is a verbose procedure giving the final relation and the need to resort to a numerical method in order to assess all the poles involved in the Laplace transform. However, the relevant simulations carried out highlight an excellent behaviour of the analytical approach: these properties can be considered attractive also considering the satisfactory overlap of the curves provided by the analytical approach with a more performing numerical model (NLC), whose excellence has been confirmed with experimental validations.

The study presented in this paper gives the possibility of investigating relevant dynamic behaviours, suggesting an efficient estimation of control and design parameters, in the frame of systems normally operating in industrial automation.

REFERENCES

Messina, A., Giannoccaro, N. I., Gentile, A., 2005. Experimenting and modelling the dynamics of pneumatic actuators controlled by the pulse width modulation (PWM) technique. *Mechatronics*, 15, 859-881.

Carducci, G., Giannoccaro, N. I., Messina, A., Rollo, G., 2006. Identification of viscous friction coefficients for a pneumatic system model using optimization methods. *Mathematics and Computers in Simulation*, 71, 385-394.

Rollo, G., 2007. *Analisi dinamiche di tipici sistemi meccanici ad attuazione pneumatica.* Ph. D. Thesis, Università del Salento, Dipartimento di Ingegneria dell'Innovazione, Lecce.

Schuder, C.B., Binder, R.C., 1959. The response of pneumatic transmission lines to step inputs. *Trans. ASME*, 81, 578-584.

Rollo, G., Messina, A., Gentile, A., 2007. Influence of geometrical parameters and operative conditions on pressure transients in pneumatic ducts. *XVIII Congresso AIMETA di Meccanica Teorica e Applicata*, MA 15, 453.

Crump, K. S., 1976. Numerical inversion of Laplace transforms using a Fourier series approximation. *Journal of the Association for Computing Machinery*, 23, 89-96.

Duffy, D. G., 1993. On the numerical inversion of Laplace transforms: comparison of three new methods on characteristic problems from application. *ACM Transaction on Mathematical Software*, 19, 333-359.

INDUCED $\ell_\infty-$ OPTIMAL GAIN-SCHEDULED FILTERING OF TAKAGI-SUGENO FUZZY SYSTEMS

Isaac Yaesh
Control Department, IMI Advanced Systems Div., P.O.B. 1044/77, Ramat–Hasharon, 47100, Israel
iyaesh@imi-israel.com

Uri Shaked
School of Electrical Engineering, Tel Aviv University, Tel Aviv, 69978, Israel
shaked@eng.tau.ac.il

Keywords: Takagi-Sugeno fuzzy systems, Polytopic uncertainties, H_∞-optimization, ℓ_∞-optimization, $\mathcal{S}$-procedure, Linear matrix inequalities.

Abstract: The problem of designing gain-scheduled filters with guaranteed induced ℓ_∞ norm for the estimation of the state-vector of finite dimensional discrete-time parameter-dependent Takagi-Sugeno Fuzzy Systems systems is considered. The design process applies a lemma which was recently derived by the authors of this paper, characterizing the induced ℓ_∞ norm by Linear Matrix Inequalities. The suggested filter has been successfully applied to a guidance motivated estimation problem, where it has been compared to an Extended Kalman Filter.

1 INTRODUCTION

The theory of optimal design of estimators for linear discrete-time systems in a state-space formulation has been first established in (Kalman, 1960). The original problem formulation assumed Gaussian white noise models for both the measurement noise and the exogenous driving process. For this case, the results of (Kalman, 1960) provided the Minimum-Mean-Square Estimator (MMSE). The Kalman filter has found since then many applications (see e.g. (Sorenson, 1985) and the references therein). Following the introduction of H_∞ control theory in (Zames, 1981), a method for designing discrete-time H_∞ optimal estimators within a deterministic framework has been developed in (Yaesh and Shaked, 1991), where the exogenous signals are of finite energy. The case where the driving signal is of finite energy (e.g. piecewise constant for a finite time) ,whereas the measurement noise is white has been recently considered in (Yaesh and Shaked, 2006). However, in some cases the minimization of the maximum absolute value of the estimation error (namely the $\ell_\infty-$norm) rather than the error energy is required where the exogenous signals are also of finite $\ell_\infty-$norm. In such cases, an induced $\ell_\infty-$norm is obtained which is often referred to as an ℓ_1 problem due to the fact that the induced-$\ell_\infty-$norm for a linear system is just the ℓ_1-norm of its impulse response and an upper-bound on the ℓ_1-norm of its transfer function (see (Dahleh and Pearson, 1987)).

In the present paper, the problem of discrete-time optimal state-estimation in the minimum induced $\ell_\infty-$norm sense is considered for a class of Takagi-Sugeno fuzzy systems. The plant model for the systems considered, is described by a collection of 'sample' finite-dimensional linear-time-invariant plants which possess the same structure but differ in their parameters. All possible plant models are then assumed to be convex combinations of these specific plant models (namely a polytopic system where the 'sample' plant models are denoted as its vertices (Boyd et al., 1994)). The solution of the estimation problem is characterized by LMIs (Linear Matrix Inequalities) based on the quadratic stability assumption. We note that more recent developments of (Geromel et al., 2000) include gain scheduled filter synthesis for the cases of linear (Tuan et al., 2001) and nonlinear (Hoang et al., 2003) dependence on the parameters with parameter dependent Lyapunov functions. We also note that in (Salcedo and Martinez, 2008) related results appear where the continuous-time fuzzy output feedback and filtering were considered in parallel to the discrete-time results of the present paper.

The paper is organized as follows. In Section 2, the problem is formulated and a key lemma characterizing the induced $\ell_\infty-$norm in terms of LMIs is presented. In Section 3 the filter design inequalities are obtained. Section 4 considers a numerical example dealing with a robust gain scheduled tracking problem. Finally, Section 5 brings some concluding remarks.

Notation: Throughout the note the superscript 'T' stands for matrix transposition, $\mathcal{R}^n$ denotes the n dimensional Euclidean space, $\mathcal{R}^{n\times m}$ is the set of all $n \times m$ real matrices, and the notation $P > 0$, for $P \in \mathcal{R}^{n\times n}$ means that P is symmetric and positive definite. The space of square summable functions over $[0 \quad \infty]$ is denoted by $l_2[0 \quad \infty]$, and $||.||_2$ stands for the standard l_2-norm,$||u||_2 = (\Sigma_{k=0}^{\infty} u_k^T u_k)^{1/2}$. We also use $||.||_\infty$ for the l_∞-norm namely, $||u||_\infty^2 = sup_k(u_k^T u_k)$. The convex hull of a and b is denoted by $\mathcal{C}o\{a,\ b\}$, I_n is the unit matrix of order n, and $0_{n,m}$ is the $n \times m$ zero matrix and $I_{m,n}$ is a version of I_n with last $n-m$ rows omitted.

2 PROBLEM FORMULATION AND PRELIMINARIES

We consider the following linear system:

$$\begin{array}{l} x(k+1) = A(k)x(k) + Bw(k), \quad x(0) = x_0 \\ y(k) = C(k)x(k) + Dw(k) \\ z(k) = L(k)x(k) \end{array} \tag{1}$$

where $x \in \mathcal{R}^n$ is the system states, $y \in \mathcal{R}^r$ is the measurement, $w \in \mathcal{R}^q$ includes the driving process and the measurement noise signals and it is assumed to have bounded $\ell_\infty-$norm. The sequence $z \in \mathcal{R}^m$ is the state combination to be estimated and A, B, C, D and L are matrices of the appropriate dimensions.

We assume that the system parameters lie within the following polytope

$$\Omega := \left[\begin{array}{ccccc} A & B & C & D & L \end{array}\right] \tag{2}$$

which is described by its vertices. That is, for

$$\Omega_i := \left[\begin{array}{ccccc} A_i & B_i & C_i & D_i & L_i \end{array}\right] \tag{3}$$

we have

$$\Omega = \mathcal{C}o\{\Omega_1, \Omega_2, ..., \Omega_N\} \tag{4}$$

where N is the number of vertices. In other words:

$$\Omega = \sum_{i=1}^{N} \Omega_i f_i \quad , \sum_{i=1}^{N} f_i = 1 \quad , f_i \geq 0. \tag{5}$$

Assuming that f_i are exactly known, the above system is just a Tagaki-Sugeno fuzzy system. To see this, one may introduce new parameters $s_i(t), i = 1, 2, ..., p$ (so called premise variables, see (Tanaka and Wang, 2001)) possibly depending on the state-vector $x(t)$, external disturbances and/or time (Tanaka and Wang, 2001) and rewrite (1) as :

IF s_1 is M_{i1} and s_2 is M_{i2} and ... s_p is M_{ip} **THEN**

$$\begin{array}{l} x(k+1) = A_i(k)x(k) + B_i w(k), \quad x(0) = x_0 \\ y(k) = C_i(k)x(k) + D_i w(k) \\ z(k) = L_i(k)x(k) \\ i = 1, 2, ..., N \end{array} \tag{6}$$

where M_{ij} is the fuzzy set and N is the number of model rules. Defining $s(t) = col\{s_1(t), s_2(t), ..., s_p(t)\}$,

$$\omega_i(s(t)) = \Pi_{j=1}^{p} M_{ij}(s_j(t))$$

and

$$f_i(s(t)) = \frac{\omega_i(t)}{\Sigma_{i=1}^{N} \omega_i(s(t))}$$

we readily get the representation of (1). We, therefore, assume indeed that the p premise scalar variables $s_i(t), i = 1, 2, ..., p$, and, consequently f_i are exactly known and consider the following filter:

$$\hat{x}(k+1) = A\hat{x}(k) + K(k)(y - C\hat{x}),\ \hat{z}(k) = L\hat{x}(k) \tag{7}$$

where the filter gain is given by the following:

$$K = \sum_{i=1}^{N} K_i f_i \tag{8}$$

and where $A = \sum_{i=1}^{N} A_i f_i$ and $L = \sum_{i=1}^{N} L_i f_i$. We will differently treat, in the sequel, the case where C is constant and the case where $C = \sum_{i=1}^{N} C_i f_i$.

Our aim is to find the filter parameters K_i so that the following induced $\ell_\infty-$norm condition is satisfied.

$$sup_{w\in\ell_\infty} ||z - \hat{z}||_\infty / ||w||_\infty < \gamma \tag{9}$$

To solve this problem we will first define another polytopic system :

$$\bar{\Omega} := \left[\begin{array}{cccc} \bar{A} & \bar{B} & \bar{C} & \bar{D} \end{array}\right] \tag{10}$$

which is described by the vertices:

$$\bar{\Omega}_i := \left[\begin{array}{cccc} \bar{A}_i & \bar{B}_i & \bar{C}_i & \bar{D}_i \end{array}\right],\ i = 1, ..., N \tag{11}$$

The system of (10)-(11) will represent, in the sequel, the dynamics of the estimation error for the system (1). The following technical lemma will be needed in order to provide convex characterization of the induced $\ell_\infty-$ norm of the estimation error system:

Lemma 1. The system

$$\begin{array}{l} \bar{x}(k+1) = \bar{A}(k)\bar{x}(k) + \bar{B}w(k), \quad x(0) = x_0 \\ z(k) = \bar{C}(k)\bar{x}(k) + \bar{D}\bar{w}(k) \end{array} \tag{12}$$

satisfies

$$sup_{w\in\ell_\infty} ||z||_\infty / ||w||_\infty < \gamma \tag{13}$$

if the following matrix inequalities are satisfied for $i = 1, 2, ..., N$:

$$\begin{bmatrix} \bar{A}_i^T P \bar{A}_i + \lambda P - P & \bar{A}_i^T P \bar{B}_i \\ \bar{B}_i^T P \bar{A}_i & -\mu I + \bar{B}_i^T P \bar{B}_i \end{bmatrix} < 0 \qquad (14)$$

and

$$\begin{bmatrix} \lambda P & 0 & \bar{C}_i^T \\ 0 & (\gamma - \mu) I & \bar{D}_i^T \\ \bar{C}_i & \bar{D}_i & \gamma I \end{bmatrix} > 0 \qquad (15)$$

so that $P > 0$, $\mu > 0$ and $\lambda < 1$.

The proof of this lemma is given in (Shaked and Yaesh, 2007) and is also provided, for the sake of completeness, in Appendix A.

Remark. Note that (14) can be written, using Schur complements ((Boyd et al., 1994)), as follows:

$$\begin{bmatrix} P - \lambda P & 0 & \bar{A}_i^T P \\ 0 & \mu I & \bar{B}_i^T P \\ P\bar{A}_i & P\bar{B}_i & P \end{bmatrix} > 0 \qquad (16)$$

or equivalently as

$$\begin{bmatrix} P - \lambda P & 0 & \bar{A}_i^T \\ 0 & \mu I & \bar{B}_i^T \\ \bar{A}_i & \bar{B}_i & P^{-1} \end{bmatrix} > 0 \qquad (17)$$

The fact that the inequality (16) is affine in $\bar{A}_i$ and $\bar{B}_i$ will be utilized in the sequel to obtain convex characterization (i.e. in LMI form) of the filter parameters K_i.

3 GAIN SCHEDULED FILTERING

Defining the state estimation error to be:

$$e(k) = x(k) - \hat{x}(k) \qquad (18)$$

we readily have for the case where f_i are available for the estimation process, that

$$e(k+1) = (A - K(k)C)e(k) + (B - K(k)D)w(k) \qquad (19)$$

and

$$z(k) - \hat{z}(k) = Le(k) \qquad (20)$$

We substitute $\bar{A}_i = A_i - K_iC$, $\bar{B}_i = B_i - K_iD$ and $\bar{C} = L_i$ in (14) and (15) where we restrict our attention to the case where C and D are not vertex dependent (i.e. $C_i = C, D_i = D, i = 1, 2, ..., N$). In this case, we define $Y_i = PK_i$ and readily obtain from (16) and (15) that

$$\begin{bmatrix} P - \lambda P & 0 & A_i^T P - C^T Y_i^T \\ 0 & \mu I & B_i^T P - D^T Y_i^T \\ PA_i - Y_iC & PB_i - Y_iD & P \end{bmatrix} > 0 \qquad (21)$$

and

$$\begin{bmatrix} \lambda P & 0 & L_i^T \\ 0 & (\gamma - \mu) I & 0 \\ L_i & 0 & \gamma I \end{bmatrix} > 0, \quad \lambda < 1 \qquad (22)$$

We, therefore, obtain the following result:

Theorem 1. Consider the estimator of (12) for the system of (1) with $C_i = C, D_i = D, i = 1, 2, ..., N$. The estimation error satisfies (9) if (21) and (22) are satisfied for $i = 1, 2, ..., N$ so that $P > 0$, $\mu > 0$ and $\lambda < 1$.

We next address the problem where C and D are vertex dependent. To this end we consider a version of Lemma 1 which can be written in terms of Ω rather than Ω_i, namely we replace (16) and (14) by:

$$\begin{bmatrix} P - \lambda P & 0 & \bar{A}^T P \\ 0 & \mu I & \bar{B}^T P \\ P\bar{A} & P\bar{B} & P \end{bmatrix} > 0 \qquad (23)$$

and

$$\begin{bmatrix} \lambda P & 0 & \bar{C}^T \\ 0 & (\gamma - \mu) I & \bar{D}^T \\ \bar{C} & \bar{D} & \gamma I \end{bmatrix} > 0 \qquad (24)$$

and substitute $\bar{A} = \sum_{i,j=1}^N (A_i - K_iC_j) f_i f_j, \bar{B}_i = \sum_{i,j=1}^N (B_i - K_iD_j) f_i f_j$ and $\bar{C} = \sum_{i=1}^N L_i f_i$. We obtain defining $Y_i = PK_i$:

$$\sum_{i,j=1}^{N} G_{ij} f_i f_j > 0 \qquad (25)$$

where

$$G_{ij} := \begin{bmatrix} P - \lambda P & 0 & A_i^T P - C_j^T Y_i^T \\ 0 & \mu I & B_i^T P - D_j^T Y_i^T \\ PA_i - Y_iC_j & PB_i - Y_iD_j & P \end{bmatrix} \qquad (26)$$

and

$$\sum_{i=1}^{N} \begin{bmatrix} \lambda P & 0 & L_i^T \\ 0 & (\gamma - \mu) I & 0 \\ L_i & 0 & \gamma I \end{bmatrix} f_i > 0 \qquad (27)$$

Since, however (see (Tanaka and Wang, 2001)) equation (25) can be also written as

$$\sum_{i,j=1}^{N} G_{ij} f_i f_j = \sum_{i=1}^{N} G_{ii} f_i^2 + 2 \sum_{i=1}^{N} \sum_{i<j} \frac{G_{ij} + G_{ji}}{2} f_i f_j \qquad (28)$$

Defining a simple transformation of the convex coordinates f_k so that for $k = 1, 2, ..., N$ we set $h_k = f_k^2$ where as the remaining h_k for $k = N+1, N+2, ..., N + \frac{N(N-1)}{2}$ are defined by $h_k = 2 f_i f_j,\ j = 1, 2..., N, i < j$. Since obviously $\sum_{k=1}^{N+\frac{N(N-1)}{2}} h_k = 1$ where $h_k \geq 0$ they can serve as convex coordinates. We, therefore, define the following LMIs inspired by (Tanaka and Wang, 2001),

$$G_{ii} > 0, i = 1, 2, ...N \text{ and } G_{ij} + G_{ji} > 0, i < j \qquad (29)$$

and obtain the following result:

Theorem 2. Consider the estimator (12) for the system (1). The estimation error satisfies (9) if (29) and (22) for $i = 1,2,...,N$ are satisfied so that $P > 0$, $\mu > 0$ and $\lambda < 1$.

The solution offered above, for the case where C and D are uncertain and are known to reside in a given polytope, seeks a single matrix P that solves the LMIs for $\frac{N(N+1)}{2}$ vertices, instead of the N vertices that were solved for in the case of known C and D. A solution for such large number of vertices by a single P entails a significant overdesign. Even the relaxation offered by e.g. (Shaked, 2003) to reduce the overdesign by allowing different $P_i, i = 1,2,..., \frac{N(N+1)}{2}$ for the $\frac{N(N+1)}{2}$ vertices still suffers from a considerable conservatism. Moreover, the computational complexity of the solution also rapidly increases as a function of the number of vertices.

In many cases, C resides in some uncertainty polytope, while D is fixed and known. In such a case, an alternative way to deal with the problem is to define $\xi(k) = col\{x(k),\ y(k)\}$ and $\tilde{w}(k) = col\{w(k),\ w(k+1)\}$so that the augmented system becomes:

$$\begin{aligned}
&\xi(k+1) = \tilde{A}(k)x(k) + \tilde{B}\tilde{w}(k)\\
&y(k) = \tilde{C}(k)\xi(k) + \tilde{D}\tilde{w}(k)\\
&z(k) = \tilde{L}(k)\xi(k)
\end{aligned}$$

where

$$\tilde{A} = \begin{bmatrix} A & 0 \\ CA & 0 \end{bmatrix} \quad \tilde{B} = \begin{bmatrix} B & 0 \\ CB & D \end{bmatrix}, \quad \tilde{C} = \begin{bmatrix} 0 & I_r \end{bmatrix},$$

$$\tilde{L} = \begin{bmatrix} L & 0 \end{bmatrix}, \text{ and } \tilde{D} = \begin{bmatrix} D & 0 \end{bmatrix} \tag{30}$$

In (30) the uncertainties appear in $\bar{A}$ and $\bar{B}$ only and, therefore, Theorem 1 above may be invoked. We, therefore, obtain the following result which offers reduced conservatism with respect the corresponding continuous-time results of (Salcedo and Martinez, 2008):

Theorem 3. Consider the estimator of (12) for the system of (1) for $D_i = D, i = 1,2,...,N$. The estimation error satisfies (9) with γ replaced by $\sqrt{2}\gamma$ if (21) and (22) are satisfied for $i = 1,2,...,N$ so that $P > 0$, $\mu > 0$ and $\lambda < 1$ with A,B,C,L replaced by $\tilde{A},\tilde{B},\tilde{C},\tilde{L}$ of (30).

4 EXAMPLE

We consider the dynamic model of guidance in a plane:

$$\begin{aligned}
\dot{\tilde{x}} &= \nu cos(\tilde{\psi}) + w_1\\
\dot{\tilde{y}} &= \nu sin(\tilde{\psi}) + w_2\\
\dot{\tilde{\psi}} &= \tilde{\phi}\\
\dot{\phi} &= -\phi/\tau + u/\tau
\end{aligned} \tag{31}$$

where $\tilde{x}$ and $\tilde{y}$ are the first two coordinates of a flight vehicle cruising in a constant altitude, in a local level north-east-down system, $\tilde{\psi}$ is the vehicle body angle with respect to the north (i.e. azimuth angle) and ϕ is the vehicle's roll angle assumed to be governed by a first-order low-pass filter dynamics having a time-constant of τ seconds, driven by the roll-angle command u. The wind velocities at the north and east directions respectively are denoted by w_1 and w_2 whereas ν is the true-air-speed. Our aim is to filter the noisy measurements of $\tilde{x}$, $\tilde{y}$ and ϕ and to estimate $\tilde{\psi}$. Defining,

$$x = col\{\tilde{x}, \tilde{y}, \nu sin(\tilde{\psi}), \nu cos(\tilde{\psi}), \phi\}$$

the measurements vector which consists of noisy measurements of the position components $\tilde{x}$ and $\tilde{y}$ and the roll angle ϕ is given by

$$y = Cx + R^{1/2}v$$

where v is the measurement noise which is taken in the simulations in the sequel as a 3-vector of zero-mean unity variance white noise sequences but for all practical purposes is assumed to be $v \in \ell_\infty$. The noise level is set by

$$R = diag\{25, 25, 0.1\}$$

and the measurement matrix is

$$C = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Note that

$$\begin{aligned}
\dot{x}_1 &= x_4\\
\dot{x}_2 &= x_3\\
\dot{x}_3 &= x_4x_5\\
\dot{x}_4 &= -x_3x_5\\
\dot{x}_5 &= -x_5/\tau + u/\tau
\end{aligned} \tag{32}$$

namely we have a bilinear system rather than a linear one. Following (Tanaka and Wang, 2001) with a series of simple manipulations, this system can be represented as a Takagi-Sugeno fuzzy system, namely as a convex combination of linear systems where the convex coordinates are online measured. To achieve such a representation we recall that $x_5 = \phi$ is measured on line, and define $s_1 = x_5$ while neglecting the small enough noise in measuring ϕ. The validity of the latter assumption will be verified in the sequel by the estimation quality we will obtain. Assuming $x_5 \in [-\phi_{max}, \phi_{max}]$ we define $f_1 = \frac{s_1 - s_{1,min}}{s_{1,max} - s_{1,min}} =$

$\frac{x_5+\phi_{max}}{2\phi_{max}}$, $f_2 = 1 - f_1$, $\alpha_1 = \phi_{max}$ and $\alpha_2 = -\phi_{max}$. We readily see that $s_1 = \phi_{max} f_1 - \phi_{max} f_2 := \alpha_1 f_1 + \alpha_2 f_2$. We note then that the system is then governed by $\dot{\xi} = A_c(\xi)\xi + B_c w$ where $\xi = col\{x_1, x_2, x_3, x_4, x_5\}$ and

$$A_c = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & s_1 & 0 \\ 0 & 0 & -s_1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1/\tau \end{bmatrix}$$

Therefore, $A_c = A_{c,1}\alpha_1 + A_{c,2}\alpha_2$ where $A_{c,1}$ is obtained from A_c by replacing s_1 by α_1 and $A_{c,2}$ is similarly obtained from A_c by replacing s_1 by α_2. We also define $w = col\{w_1, w_2, \nu_1, \nu_2, \nu_3\}$ and complete the remaining matrices needed for the representation of our problem (1)-(3) by applying a zero-order-hold discrete-time equivalent of our continuous-time system, where we have chosen a sampling time of $h = 0.02$. Due to the small enough h we have chosen, we have $e^{Ah} = I + Ah + O(h^2)$ and we, therefore, readily obtain that the system is governed by (1) and (3)-(5) where $A = A_1\alpha_1 + A_2\alpha_2 + O(h^2)$ where

$$A_1 = \begin{bmatrix} 1.0000 & -0.0002 & 0.0200 & 0 & \\ 0 & 1.0000 & 0.0200 & 0.0002 & 0 \\ 0 & 0 & 0.9998 & 0.0209 & 0 \\ 0 & 0 & -0.0209 & 0.9998 & 0 \\ 0 & 0 & 0 & 0 & 0.9231 \end{bmatrix}$$

$$A_2 = \begin{bmatrix} 1.0000 & 0 & 0.0002 & 0.0200 & 0 \\ 0 & 1.0000 & 0.0200 & -0.0002 & 0 \\ 0 & 0 & 0.9998 & -0.0209 & 0 \\ 0 & 0 & 0.0209 & 0.9998 & 0 \\ 0 & 0 & 0 & 0 & 0.9231 \end{bmatrix}$$

$$B_1 = B_2 = B = \begin{bmatrix} 0.2000 & 0 & 0 & 0 & 0 \\ 0 & 0.2000 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

and

$$D_1 = D_2 = D = \begin{bmatrix} 0 & 0 & 2.2361 & 0 & 0 \\ 0 & 0 & 0 & 2.2361 & 0 \\ 0 & 0 & 0 & 0 & 0.1000 \end{bmatrix}$$

We note at his point that, in order to minimize the design conservatism stemming from the quadratic stability assumption, we applied a parameter dependent Lyapunov function (Boyd et al., 1994), $max(x^T P_1 x, x^T P_2 x)$. Minimization of γ subject the the LMis that are obtained with this function to replace (21) and (22) (see Appendix B), using *fminsearch.m* from the optimization toolbox of $MATLAB^{TM}$ and (Lagarias et al., 1998) to search λ_i, $i = 1,2,3,4$, ρ_1, $\rho 2_2$, θ_1 and θ_2, has resulted in $\gamma = \gamma_0 = 10.2732$ and $\lambda = 2.41 \times 10^{-7}$. The following gain matrices K_1 and K_2 have been obtained for $\gamma = \gamma_0$ were obtained:

$$K_1 = \begin{bmatrix} 0.7574 & -0.0007 & 0.0000 \\ -0.0018 & 0.7593 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & -0.0000 & 0.0000 \\ -0.0000 & -0.0000 & 0.0003 \end{bmatrix}$$

$$K_2 = \begin{bmatrix} 0.7558 & -0.0024 & 0.0000 \\ -0.0021 & 0.7613 & -0.0000 \\ -0.0000 & 0.0000 & -0.0000 \\ 0.0000 & 0.0000 & 0.0000 \\ -0.0000 & 0.0000 & 0.0003 \end{bmatrix}$$

This ℓ_∞ filter will be compared to an Extended Kalman Filter (EKF, see (Jazwinsky, 1970)) based on the nonlinear model of (31). Note that higher complexity filters such as the particle filter (e.g. (Oshman and Carmi, 2006) are out the scope of the present paper. For the simulations we take $\nu = 100m/s$ and try to control the vehicle to follow a constant command at $y = 5m$, in spite of a wind step at w_2 of $10m/s$. The estimation results are used to control the vehicle, using the simple law $u = -\begin{bmatrix} 0.0200 & 4.0000 \end{bmatrix} \begin{bmatrix} \hat{y}-5 & \hat{\psi} \end{bmatrix}^T$ where all components of the initial state-vector are taken as zero, besides $y_0 = 20m$. The EKF and the ℓ_∞ estimated $\tilde{x} - \tilde{y}$ trajectory results are compared in Fig. 1 to the true trajectory. One can notice the bias in the EKF estimate. In Fig. 2, the true $\tilde{\psi}$ and the estimated values for $\tilde{\psi}$ that are obtained by using the EKF and the ℓ_∞ filter are depicted. We clearly see in this figure that the ℓ_∞ filter outperforms the EKF which assumes a white noise w_2 but leads to a bias when w_2 has a bias. In contrast, the $\tilde{\psi}$ estimate of the ℓ_∞ filter is barely separable from the true values. Moreover, the ℓ_∞ filter does not require the on-line numerical solution of a Riccati equation of order 4 and the gains are obtained there by a simple convex interpolation on K_1 and K_2. The fact that K_1 and K_2 are close to each other is somewhat surprising. Our experience shows that for a larger γ (i.e. suboptimal values), a larger $||K_1 - K_2||$ is obtained.

5 CONCLUSIONS

The problem of designing robust gain-scheduled filters with guaranteed induced $\ell_\infty-$norm has been considered. The solution has been derived using a recently developed bounded-real-lemma like condition for bounding the induced ℓ_∞ norm of a system. This result has been applied to derive the robust induced $\ell_\infty-$filter (or equivalently robust ℓ_1 filter) in terms of LMIs. These LMIs have been applied to a guidance motivated estimation example. In this example, the

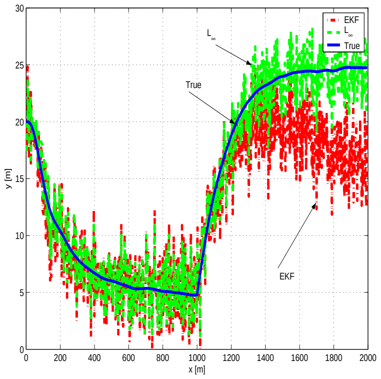


Figure 1: True, Extended Kalman Filter and ℓ_∞-Filter Estimated Trajectories - $\tilde{x}(t)$ versus $\tilde{y}(t)$.

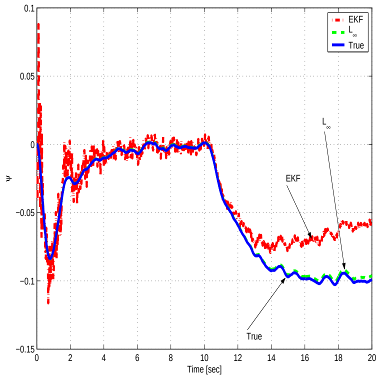


Figure 2: True, Extended Kalman Filter and ℓ_∞-Filter Estimated Azimuth Angle - $\tilde{\psi}$ versus t.

superiority of the induced ℓ_∞ filter over the Extended Kalman Filter has been demonstrated, both in terms of performance and simplicity of implementation.

REFERENCES

Abedor, J., Nagpal, K., and Poolla, K. (1996). A linear matrix inequality approach to peak-to-peak gain minimization. *International J. of Robust and Nonlinear Control*, 6:899–927.

Boyd, S., El-Ghaoui, L., Feron, L., and Balakrishnan, V. (1994). *Linear matrix inequalities in system and control theory*. SIAM.

Dahleh, M. A. and Pearson, J. (1987). ℓ_1-optimal feedback controllers for mimo discrete-time systems. *IEEE Trans. on Automat. Contr.*, AC-32:314–322.

Geromel, J., Bernussou, J., Garcia, G., and de Oliviera, M. (2000). h_2 and h_∞ robust filtering for discrete-time linear systems. *SIAM Journal of Control and Optimization*, 38:1353–1368.

Hoang, N., Tuan, H., Apkarian, P., and Hosoe, S. (2003). Robust filtering for uncertain nonlinearly parameterized plants. *IEEE Trans. on Signal Processing*, 51:1806–1815.

Jazwinsky, A. H. (1970). *Stochastic Processes and Filtering Theory*. Academic Press, New-York.

Kalman, R. (1960). A new approach to linear filtering and prediction problems. *Transactions ASME, Journal of Basic Engineering*, 82d:35–45.

Lagarias, J., Reeds, J. A., Wright, M. H., and Wright, P. E. (1998). Convergence properties of the nelder-mead simplex method in low dimensions. *SIAM Journal of Optimization*, 9:112–147.

Oshman, Y. and Carmi, A. (2006). Attitude estimation from vector obsevations usinga genetic algorithm-embedded quaternion particle filter. *Journal of Guidance, Control and Dyanmics*, 29:879–891.

Salcedo, J. V. and Martinez, M. (2008). Bibo stabilization of takagi-sugeno fuzzy systems under persistent perturbations using fuzzy output feedback controllers. *IET Control Theory and Applications*, 2:513–523.

Shaked, U. (2003). An lpd approach to robust h_2 and h_∞ static output-feedback design. *IEEE Trans. on Automat. Contr.*, 48:866–872.

Shaked, U. and Yaesh, I. (2007). Robust servo synthesis by minimization of induced ℓ_2 and ℓ_∞ norms. In *ISIE 2007, Vigo, Spain*.

Sorenson, H. (1985). *Kalman Filtering : Theory and Application*. IEEE Press.

Tanaka, K. and Wang, H. (2001). *Fuzzy Control Systems Design and Analysis - A Linear Matrix Inequality Approach*. John Wiley and Sons, Inc.

Tuan, H. D., Apkarian, P., and Nguyen, T. Q. (2001). Robust and reduced-order filtering: new lmi-based characterizations and methods. *IEEE Trans. on Signal Processing*, 40:2975–2984.

Yaesh, I. and Shaked, U. (1991). A transfer function approach to the problems of discrete-time systems h_∞ -optimal control and filtering. *IEEE Trans. on Automat. Contr.*, 36:1264–1271.

Yaesh, I. and Shaked, U. (2006). Discrete-time min-max tracking. *IEEE Trans. Aerospace and Elect. Systems*, 42:540–547.

Zames, G. (1981). Feedback and optimal sensitivity : model reference transformations, multiplicative seminorms and approximate inverses. *IEEE Trans. on Automat. Contr.*, AC-26:301–320.

APPENDIX A - PROOF OF LEMMA 1

Consider the system

$$x_{k+1} = \bar{A}x_k + \bar{B}w_k, \quad z_k = \bar{C}x_k + \bar{D}w_k$$

and define, following (Abedor et al., 1996),

$$\xi_k = x_{k+1}^T P x_{k+1} - x_k^T P x_k + \lambda x_k^T P x_k - \mu w_k^T w_k.$$

Namely,

$$\xi_k = (x_k^T \bar{A}^T + w_k^T \bar{B}^T) P (\bar{A}x_k + \bar{B}w_k) - x_k^T P x_k + \lambda x_k^T P x_k - \mu w_k^T w_k.$$

Collecting terms we have

$$\xi_k = x_k^T(\bar{A}^T P\bar{A} + \lambda P - P)x_k + x_k^T(\bar{A}^T P\bar{B})w_k$$
$$+ w_k^T(\bar{B}^T P\bar{A})x_k + w_k^T(-\mu I + \bar{B}^T P\bar{B})w_k.$$

Therefore, (14) guarantees $\xi_k < 0$ for all w_k and x_k.

Defining $\zeta_k = x_k^T P x_k$ and assuming $x_0 = 0$ and $w_k^T w_k < 1$ we have that $\zeta_{k+1} - \zeta_k + \lambda\zeta_k - \mu w_k^T w_k < 0$. Namely, $\zeta_k < \bar{\zeta}_k$ where $\bar{\zeta}_{k+1} = (1-\lambda)\bar{\zeta}_k + \mu w_k^T w_k$. However, using $\rho := 1 - \lambda$ we have

$$\bar{\zeta}_k = \sum_{j=0}^{k-1} \rho^{k-j-1}\mu w_j^T w_j = \rho^{k-1}\sum_{j=0}^{k-1}(\rho^{-1})^j \mu w_j^T w_j$$

$$< \rho^{k-1}\frac{(1-\rho^{-1})^k}{1-\rho^{-1}}\mu = \mu\frac{1-\rho^k}{1-\rho}. \qquad \mathbf{A.1}$$

From (15) we have using Schur complements that

$$[x^T\ w^T]\left(\begin{bmatrix} \lambda P & 0 \\ 0 & (\gamma-\mu)I \end{bmatrix} - \gamma^{-1}\begin{bmatrix} \bar{C}^T \\ \bar{D}^T \end{bmatrix}[\ \bar{C} \quad \bar{D}\ ]\right)\begin{bmatrix} x \\ w \end{bmatrix} > 0$$

Namely,

$$z_k^T z_k < \gamma[(\gamma-\mu)w_k^T w_k + \lambda x_k^T P x_k] < \gamma[(\gamma-\mu) + \lambda\bar{\zeta}_k] \ \mathbf{A.2}$$

Substituting A.1 we readily see that

$$z_k^T z_k < \gamma[(\gamma-\mu) + (1-\rho)\mu\frac{1-\rho^k}{1-\rho}] = \gamma[\gamma - \mu + \mu - \mu\rho^k].$$

Since $0 < \rho < 1$ we obtain that

$$z_k^T z_k < \gamma[\gamma - \mu + \mu] = \gamma^2.$$

APPENDIX B - PARAMETER DEPENDENT RESULTS

In order to reduce conservatism, we replace in the proof of Lemma 1 in Appendix A, the parameter-independent Lyapunov function $V(x,P) = x_k^T P x_k$ by the parameter-dependent Lyapunov function ((Boyd et al., 1994)) $V(x,P_1,P_2) = max(x_k^T P_1 x_k, x_k^T P_2 x_k)$. To ensure $V(x_k,P_1,P_2) > 0$ we have to satisfy $x_k^T P_1 x_k > 0$ whenever $x_k^T P_1 x_k > x_k^T P_2 x_k$ and $x_k^T P_2 x_k > 0$ whenever $x_k^T P_1 x_k < x_k^T P_2 x_k$. Applying the S-procedure (Boyd et al., 1994) we readily obtain that a sufficient condition or these requirements to hold, is the existence of constants $\rho_1 > 0$ and $\rho_2 > 0$ so that

$$P_1 - \rho_1(P_1 - P_2) > 0 \text{ and } P_2 - \rho_2(P_2 - P_1) > 0.$$

We also require that if $x_k^T P_1 x > x_k^T P_2 x_k$ then

$$\xi_k^{\{1\},\bar{A}} := x_k^T(\bar{A}^T P_1\bar{A} + \lambda P_1 - P_1)x_k + x_k^T(\bar{A}^T P_1\bar{B})w_k$$
$$+ w_k^T(\bar{B}^T P_1\bar{A})x_k + w_k^T(-\mu I + \bar{B}^T P_1\bar{B})w_k < 0,$$

and if $x_k^T P_2 x > x_k^T P_1 x_k$ then

$$\xi_k^{\{2\},\bar{A}} := x_k^T(\bar{A}^T P_2\bar{A} + \lambda P_2 - P_2)x_k + x_k^T(\bar{A}^T P_2\bar{B})w_k$$
$$+ w_k^T(\bar{B}^T P_2\bar{A})x_k + w_k^T(-\mu I + \bar{B}^T P_2\bar{B})w_k < 0.$$

Since these conditions are required to be satisfied throughout the polytope, we readily obtain, using again the S-procedure, that in addition to the constant $\lambda > 0$, the existence of six additional constants $\lambda_i > 0$, $i = 1,2,3,4$, $\theta_1 > 0$, $\theta_2 > 0$ establishes a sufficient condition for the above inequalities to hold, if

$$-\xi_k^{\{1\},\bar{A}_i} - \lambda_1(P_1 - P_2) > 0, i = 1,2$$

and

$$-\xi_k^{\{1\},\bar{A}_i} - \lambda_2(P_2 - P_1) > 0, i = 1,2.$$

Following the lines of proof of Theorem 1 above, we readily obtain the following LMIs for $i = 1,2$ to replace (21) and (22):

$$\begin{bmatrix} P_1 - \lambda P - \lambda_i(P_1 - P_2) & 0 & A_i^T P_1 - C^T Y_i^T \\ 0 & \mu I & B_i^T P_1 - D^T Y_i^T \\ P_1 A_i - Y_i C & P_1 B_i - Y_i D & P_1 \end{bmatrix} > 0,$$

$$\begin{bmatrix} \lambda P_1 - \theta_1(P_1 - P_2) & 0 & L_i^T \\ 0 & (\gamma-\mu)I & 0 \\ L_i & 0 & \gamma I \end{bmatrix} > 0, \quad \lambda < 1$$

and

$$\begin{bmatrix} P_2 - \lambda P - \lambda_{i+2}(P_2 - P_1) & 0 & A_i^T P_2 - C^T Y_i^T \\ 0 & \mu I & B_i^T P_2 - D^T Y_i^T \\ P_2 A_i - Y_i C & P_2 B_i - Y_i D & P_2 \end{bmatrix} > 0,$$

$$\begin{bmatrix} \lambda P_2 - \theta_2(P_2 - P_1) & 0 & L_i^T \\ 0 & (\gamma-\mu)I & 0 \\ L_i & 0 & \gamma I \end{bmatrix} > 0, \quad \lambda < 1.$$

We note that we have also replaced (A.2) with:

$$z_k^T z_k < \gamma[(\gamma-\mu)w_k^T w_k + \lambda \times max(x_k^T P_1 x_k, x_k^T P_2 x_k)].$$

Namely, if $x^T P_1 x > x^T P_2 x$ we require

$$z_k^T z_k < \gamma[(\gamma-\mu)w_k^T w_k + \lambda x_k^T P_1 x],$$

whereas if $x^T P_2 x > x^T P_1 x$ we require

$$z_k^T z_k < \gamma[(\gamma-\mu)w_k^T w_k + \lambda x_k^T P_2 x].$$

Using again the $\mathcal{S}$-Procedure with additional tuning constants $\theta_1 > 0$ and $\theta_2 > 0$, which add up to the previously introduced 7 tuning constants $\rho_1 > 0$, $\rho_2 > 0$, $\lambda > 0$ and $\lambda_i > 0, i = 1,2,3,4$ the above results are obtained. Note that a similar approach can be applied also on the continuous-time results of (Salcedo and Martinez, 2008) to reduce conservatism.

A NEW METHOD OF TUNING THREE TERM CONTROLLERS FOR DEAD-TIME PROCESSES WITH A NEGATIVE/POSITIVE ZERO

K. G. Arvanitis, G. D. Pasgianos
Department of Natural Resources Management and Agricultural Engineering
Agricultural University of Athens, 75 Iera Odos Str., 11855, Athens, Greece
karvan@aua.gr, pasgianos@geomations.com

A. K. Boglou
Technology Education Institute of Kavala, 65404, Agios Loukas, Kavala, Greece
akbogl@teikav.edu.gr

N. K. Bekiaris-Liberis
Department of Mechanical and Aerospace Engineering, University of California
San Diego, La Jolla, CA 92093-0411, U.S.A.
nbekiari@ucsd.edu

Keywords: Process Control, PID Control, Dead-Time Processes, Process Zeros.

Abstract: The use of the Pseudo-Derivative Feedback (PDF) structure in the control of stable or unstable dead-time processes with a negative or a positive zero is investigated. A simple direct synthesis method for tuning the PDF controller is presented. Moreover, a modification of the proposed method, which ensures its applicability in the case of large overshoot response processes with dead time, is also presented. The PDF control structure and the proposed tuning method ensure smooth closed-loop response to set-point changes, fast regulatory control and sufficient robustness against parametric uncertainty. Simulation results show that, in most cases, the proposed method is as efficient as the best of the most recent PID controller tuning methods for dead-time processes with negative/positive zeros, while its simplicity in deriving the controller settings is a plus point over existing PID controller tuning formulae.

1 INTRODUCTION

In the process industry, stable second order dead-time models as well as second order dead-time models with one right-half-plane pole are frequently used to adequately describe numerous processes for the purpose of designing controllers. However, these types of process models are inadequate in the case where a process controlled variable encounters two (or more) competing dynamic effects with different time constants from the same manipulated variable (Waller and Nygardas, 1975). This composite dynamics results to a process behaviour that exhibits an inverse response or a large overshoot response. Inverse response or large overshoot response is portrayed by the presence of one (or an odd number of) positive or negative zeros, respectively, in the process models, and they can cause, together with the process dead-times, serious problems in designing and tuning simple controllers for the process under consideration.

Inverse response second order dead-time process models (SODT-IR) are used to represent the dynamics of several chemical processes (like level control loops in distillation columns and temperature control loops in chemical reactors), as well as the dynamics of PWM based DC-DC boost converters in industrial electronics. In the extant literature, there is a number of studies regarding the design and tuning of three-term controllers for SOPD-IR processes (Waller and Nygardas, 1975; Scali and Rachid, 1998; Zhang *et al*, 2000; Luyben, 2000; Chien *et al,* 2003; Padma Sree and Chidambaram, 2004; Chen *et al,* 2005; Chen *et al,* 2006). In particular, Waller and Nygardas (1975) presented an empirical tuning of PID controllers based on the Ziegler-Nichols method for SOPDT-IR processes. In Scali and Rachid (1998) and Zhang *et al* (2000), analytical design methods based on the Internal Model Control framework and the H_∞ control theory, have been proposed for inverse response processes without time delay. In Luyben (2000), an empirical method that gives large overshoot and oscillatory response has been proposed to design PI controllers for SODT-IR processes. In Chien *et al* (2003), a direct synthesis tuning method is presented to tune PID controllers for both under-damped and over-damped SODT-IR processes. In Chen *et al* (2005), an analytical PID controller design for SOD-IR processes is derived based on conventional unity feedback control. In Chen *et al* (2006), an analytical design scheme based on IMC theory has been proposed to control SODT-IR processes. Finally, in Padma Sree and Chidambaram (2004), a method of tuning set-point weighted PID controllers for unstable SODT processes with a positive or a negative zero is presented. This method is based on appropriately equating coefficients of like powers of s in the numerator and the denominator of the closed-loop transfer function.

In contrast, controller tuning for large overshoot response dead-time processes have received less attention in the past, although they used to model several physical phenomena, like blending processes, mixing processes in distillation columns and temperature of heat exchangers (see Chien *et al* (2003), for details). In Chang *et al* (1997) a tuning method of controllers in first order lead-lag form has been proposed for such processes. Furthermore in Chien *et al* (2003), a direct synthesis tuning method is presented in order to tune PID controllers for both under-damped and over-damped large overshoot response processes.

The present paper investigates some aspects of the controller configuration proposed by Phelan (1978), and called the “pseudo-derivative feedback controller” (PDF), which is put forward here as an alternative means of tuning three-term controllers for stable or unstable dead time processes with a negative or positive zero. The aim of the paper is to propose a set of tuning rules for the PDF controller when it is applied to such processes. The proposed method is a direct synthesis tuning method and it is based on the manipulation of the closed loop transfer function through appropriate approximations of the dead-time term in the denominator of the closed loop transfer function as well as appropriate selection of the derivative gain, in order to obtain a second order dead-time closed-loop system. On the basis of this method the settings of the PDF controller are obtained in terms of two adjustable parameters, one of which can further be appropriately selected in order to achieve a desired damping ratio for the closed-loop system, while the other is free to designer and can be selected in order to enhance the obtained regulatory control performance. Moreover, an appropriate modification of the proposed method, that makes it applicable in the case of large overshoot response processes with dead time, is also presented. For assessment of the effectiveness of the proposed tuning method and in order to provide a comparison with existing tuning methods, a series of simulation examples are presented. Simulation results verify that the PDF control structure and the proposed direct synthesis tuning method ensure smooth closed-loop response to set-point changes, fast regulatory control and sufficient robustness in case of model mismatch.

2 THE PSEUDO-DERIVATIVE FEEDBACK CONTROLLER

The Pseudo-Derivative Feedback (PDF) controller has first been proposed by Phelan (1978), and its general feedback configuration is shown in Figure 1. The transfer function $G_{CL}(s)$ of the closed loop system is given by

$$G_{CL}(s) = \frac{K_I G_P(s)}{s + \left(K_{D,n-1}s^n + \ldots + K_{D,1}s^2 + K_{D,0}s + K_I\right)G_P(s)} \quad (1)$$

The PDF controller is essentially a variation of the conventional PID controller. In contrast to the PID controller, the PDF controller does not contribute to closed-loop zeros, and hence it is expected that it will not render worst the overshoot of the closed-loop response. The two configurations differ in the way they react to set-point changes (as it can be easily checked, they are equivalent for load or disturbance changes). The PID controller often has an abrupt response to a step change because the step is amplified and transmitted directly to the feedback control element and downstream blocks. This can induce a significant overshoot in the response that is unrelated to the closed loop system damping. For this reason, it is a common practice to ramp or filter the set-point. The PDF structure avoids this because naturally ramps the controller effort, since it internalizes the pre-filter that one would apply to cancel any closed-loop zeros introduced in the PI/PID control configuration.

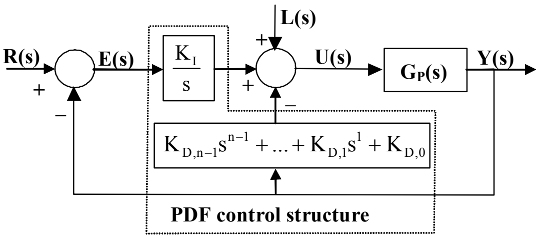


Figure 1: The general PDF control structure.

In the present paper, we focus our attention on the specific form of the general PDF control structure which contains proportional as well as a single derivative action in the feedback path (i.e. $K_{D,i}$=0 for i=2,…,n-1 and K_P≠0, $K_{D,1}$≠0). We call this feedback scheme, the PD-1F control structure, in contradistinction with the PDF controller without derivative action (i.e. the controller with $K_{D,1}=0$), which is designated as the PD-0F controller. We shall next analyze its performance, in the case where the system under control is a second order process with both dead-time and a minimum or a non-minimum phase zero, which can be described by the following general transfer function model

$$G_P(s)=K(ps+1)\exp(-ds)/[(\tau_1 s+1)(\tau_2 s+q)] \quad (2)$$

where

$$p=\begin{cases}-\tau_z & \text{in the case of a positive zero}\\ \tau_z & \text{in the case of a negative zero}\end{cases},\ \tau_z>0$$

and q=1, in the case of a stable process or q=-1, in the case of an unstable process, and where, K, d, τ_z, τ_1 and τ_2, are the process gain, the dead-time, the zero's time constant and the process time constants, respectively.

To this end, observe that, equation (1), in the case of a PD-1F controller and for process models of the form (2), takes the form

$$G_{CL}(s)=\frac{KK_I[(ps+1)/(\tau_1 s+1)]\exp(-ds)}{s(\tau_2 s+q)+(KK_d s^2+KK_P s+KK_I)\dfrac{(ps+1)}{(\tau_1 s+1)}\exp(-ds)} \quad (3)$$

Relation (3) will be the starting point for the development of the tuning method that will be presented in the sequel.

3 A SIMPLE TUNING METHOD

In order to systematically present the proposed tuning method, observe that by making use of the approximation

$$(ps+1)/(\tau_1 s+1)\approx 1-(\tau_1-p)s \quad (4)$$

in the numerator of (3), and observing that

$$\exp(-ds)=\frac{\exp[-(1-\alpha)ds]}{\exp(\alpha ds)}$$

for some $\alpha\in\Re$, we obtain

$$G_{CL}(s)=\frac{[1-(\tau_1-p)s]\exp(-ds)}{s\left(\dfrac{\tau_2}{KK_I}s+\dfrac{q}{KK_I}\right)+\left(\dfrac{K_d}{K_I}s^2+\dfrac{K_P}{K_I}s+1\right)\dfrac{(ps+1)\exp[-(1-\alpha)ds]}{(\tau_1 s+1)\exp(\alpha ds)}} \quad (5)$$

Next, using the approximations,

$$(ps+1)\exp[-(1-\alpha)ds]=[p-(1-\alpha)d]s+1$$
$$(\tau_1 s+1)\exp(\alpha ds)=(\tau_1+\alpha d)s+1$$

in (5), we obtain

$$G_{CL}(s)=\frac{[1-(\tau_1-p)s]\exp(-ds)}{s\left(\dfrac{\tau_2}{KK_I}s+\dfrac{q}{KK_I}\right)+\left(\dfrac{K_d}{K_I}s^2+\dfrac{K_P}{K_I}s+1\right)\left[\dfrac{[p-(1-\alpha)d]s+1}{(\tau_1+\alpha d)s+1}\right]} \quad (6)$$

Relation (6) may further be written as

$$G_{CL}(s)=\frac{[1-(\tau_1-p)s]\exp(-ds)}{s\left(\dfrac{\tau_2}{KK_I}s+\dfrac{q}{KK_I}\right)+P(s)} \quad (7)$$

where

$$P(s)=\left[\frac{K_d}{(\tau_1+\alpha d)K_I}s+\left(\frac{K_P}{(\tau_1+\alpha d)K_I}-\frac{K_d}{(\tau_1+\alpha d)^2K_I}\right)+Q(s)\right]$$
$$\times[[p-(1-\alpha)d]s+1]$$

$$Q(s)=\frac{1-\dfrac{K_P}{(\tau_1+\alpha d)K_I}+\dfrac{K_d}{(\tau_1+\alpha d)^2K_I}}{(\tau_1+\alpha d)s+1}$$

Observe now that by selecting

$$K_d = (\tau_1 + \alpha d)K_P - (\tau_1 + \alpha d)^2 K_I \tag{8}$$

we obtain Q(s)=0 and

$$P(s) = \left[\left(\frac{K_p}{K_I} - \tau_1 - \alpha d\right)s + 1\right]\left[\left[p - (1-\alpha)d\right]s + 1\right]$$

Therefore, relation (7) yields

$$G_{CL}(s) = \frac{\left[1-(\tau_1 - p)s\right]\exp(-ds)}{s\left(\frac{\tau_2}{KK_I}s + \frac{q}{KK_I}\right) + \left[\left(\frac{K_p}{K_I} - \tau_1 - \alpha d\right)s + 1\right]\left[\left[p-(1-\alpha)d\right]s+1\right]} \tag{9}$$

which can further be written in the form

$$G_{CL}(s) = \frac{\left[1-(\tau_1 - p)s\right]\exp(-ds)}{\lambda^2 s^2 + 2\zeta\lambda s + 1} \tag{10}$$

$$\lambda = \sqrt{\frac{\tau_2}{KK_I} - \left(\frac{K_p}{K_I} - \tau_1 - \alpha d\right)\left[(1-\alpha)d - p\right]} \tag{11}$$

$$\zeta = \frac{\frac{K_p}{K_I} - d - \tau_1 + p + \frac{q}{KK_I}}{2\sqrt{\frac{\tau_2}{KK_I} - \left(\frac{K_p}{K_I} - \tau_1 - \alpha d\right)\left[(1-\alpha)d - p\right]}} \tag{12}$$

The Routh stability criterion about (10) yields

$$K_P > (d + \tau_1 - p)K_I - \frac{q}{K} \tag{13}$$

and

$$K_P < (\tau_1 + \alpha d)K_I + \frac{\tau_2}{K\left[(1-\alpha)d - p\right]} \tag{14}$$

Therefore, as for K_P one can choose the middle value of the range given by inequalities (13) and (14). That is

$$K_P = \frac{\left[2\tau_1 + (1+\alpha)d - p\right]K_I + \frac{\tau_2 - q\left[(1-\alpha)d - p\right]}{K\left[(1-\alpha)d - p\right]}}{2} \tag{15}$$

Then, from (15), we obtain

$$\beta \mathrel{\hat{=}} \frac{K_P}{K_I} = \frac{\left[2\tau_1 + (1+\alpha)d - p\right] + \frac{\tau_2 - q\left[(1-\alpha)d - p\right]}{K_I K\left[(1-\alpha)d - p\right]}}{2} \tag{16}$$

which yields,

$$K_I = \frac{\tau_2 - q\left[(1-\alpha)d - p\right]}{K\left[(1-\alpha)d - p\right]\left[2\beta - 2\tau_1 - (1+\alpha)d + p\right]} \tag{17}$$

Therefore,

$$K_P = \frac{\beta\left[\tau_2 - q\left[(1-\alpha)d - p\right]\right]}{K\left[(1-\alpha)d - p\right]\left[2\beta - 2\tau_1 - (1+\alpha)d + p\right]} \tag{18}$$

$$K_d = \frac{\left[\beta(\tau_1 + \alpha d) - (\tau_1 + \alpha d)^2\right]\left[\tau_2 - q\left[(1-\alpha)d - p\right]\right]}{K\left[(1-\alpha)d - p\right]\left[2\beta - 2\tau_1 - (1+\alpha)d + p\right]} \tag{19}$$

Clearly, relations (17)-(19) provide the settings of the desired PD-1F controller as functions of two adjustable parameters α and β, which must be selected in order to guarantee positive controller settings (in the case where the process parameters take positive values), as well as to fulfil inequalities (13) and (14). For a pre-specified value of $\alpha \in \Re$, parameter β can be selected in order to assign a specific damping ratio ζ_{des} of the closed-loop system. Indeed, using relations (12), (17) and the definition of β, and after some trivial algebra, one can resort the following quadratic equation with regard to β,

$$A_2\beta^2 + A_1\beta + A_0 = 0 \tag{20}$$

$$\begin{aligned} A_1 = {} & 4\zeta_{des}^2\left[\frac{2}{T_2 - q\left[(1-\alpha)d - p\right]} - 1\right]\left[(1-\alpha)d - p\right] \\ & - 2\left[1 + \frac{2q\left[(1-\alpha)d - p\right]}{T_2 - q\left[(1-\alpha)d - p\right]}\right] \\ & \times\left[d + \tau_1 - q + \frac{q\left[(1-\alpha)d - p\right]\left[2\tau_1 + (1+\alpha)d - p\right]}{T_2 - q\left[(1-\alpha)d - p\right]}\right] \end{aligned} \tag{21}$$

$$A_2 = \left[1 + \frac{2q\left[(1-\alpha)d - p\right]}{T_2 - q\left[(1-\alpha)d - p\right]}\right]^2 \tag{22}$$

$$\begin{aligned} A_0 = {} & \left[d + \tau_1 - q + \frac{q\left[(1-\alpha)d - p\right]\left[2\tau_1 + (1+\alpha)d - p\right]}{\tau_2 - q\left[(1-\alpha)d - p\right]}\right]^2 \\ & - 4\zeta_{des}^2\left[\frac{\tau_2\left[2\tau_1 + (1+\alpha)d - p\right]}{\tau_2 - q\left[(1-\alpha)d - p\right]} - \tau_1 - \alpha d\right]\left[(1-\alpha)d - p\right] \end{aligned} \tag{23}$$

Then, β is chosen as the maximum real root of (20)

Clearly, the method presented above is applicable when $p=\tau_Z$ or $-\tau_Z$ and q=1 or -1. However, extensive simulations show that, in the case where $\tau_Z >> 0$ (i.e. in the case of large overshoot processes), the method provides controller settings that renders the closed-loop unstable or marginally stable. This is due to the swings of the controller output induced by the excessive derivative action. One way to avoid this problem is to filter the controller output using a first order filter of the form (see Figure 2).

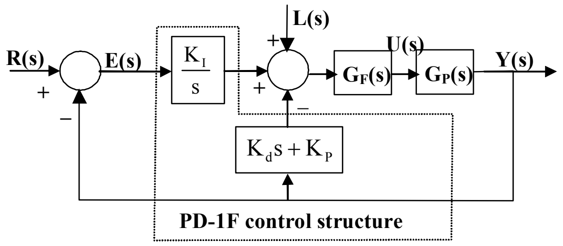


Figure 2: PD-1F control structure in the case of large overshoot response processes.

$$G_F(s) = 1/\left(\tau_F s + 1\right) \tag{24}$$

while calculating the PD-1F controller settings as suggested by relations (17)-(19). The time constant of the filter can be selected as τ_F= τ_Z.

Alternatively, one can select the controller settings according to the following method, which is a modification of the method resulting in the settings given by relations (17)-(19): In the case where the filter of the form (24) is introduced in the control loop, relation (3) is modified as

$$G_{CL}(s) = \frac{KK_I\left[\dfrac{ps+1}{(\tau_F s+1)(\tau_1 s+1)}\right]\exp(-ds)}{s(\tau_2 s+q)+\left(KK_d s^2+KK_P s+KK_I\right)\left[\dfrac{ps+1}{(\tau_F s+1)(\tau_1 s+1)}\right]\exp(-ds)}$$

Then, making use of the approximations

$$\left(ps+1\right)/\left(\tau_F s+1\right)\left(\tau_1 s+1\right) \approx 1-(\tau_1-\overline{p})s$$

$$\exp(-ds) = \frac{\exp\left[-\left(1-\alpha\right)ds\right]}{\exp(\alpha ds)}$$

$$\left(\frac{ps+1}{\tau_F s+1}\right)\exp\left[-(1-\alpha)ds\right] = \left[\overline{p}-(1-\alpha)d\right]s+1$$

$$\left(\tau_1 s+1\right)\exp(\alpha ds) = \left(\tau_1+\alpha d\right)s+1$$

where, $\overline{p} = p - \tau_F$, we finally obtain

$$G_{CL}(s) = \frac{\left[1-\left(\tau_1-\overline{p}\right)s\right]\exp(-ds)}{s\left(\dfrac{\tau_2}{KK_I}s+\dfrac{q}{KK_I}\right)+\left[\left(\dfrac{K_p}{K_I}-\tau_1-\alpha d\right)s+1\right]\left[\left[\overline{p}-(1-\alpha)d\right]s+1\right]} \tag{25}$$

It is now obvious that relation (25) is similar to relation (9) when p is replaced by $\overline{p} = p - \tau_F$. Therefore, following an argument similar to that used above to produce relations (17)-(19), we may easily conclude that, in the present case

$$K_I = \frac{\tau_2 - q\left[(1-\alpha)d-\overline{p}\right]}{K\left[(1-\alpha)d-\overline{p}\right]\left[2\beta-2\tau_1-\left(1+\alpha\right)d+\overline{p}\right]}$$

$$K_P = \frac{\beta\left[\tau_2 - q\left[(1-\alpha)d-\overline{p}\right]\right]}{K\left[(1-\alpha)d-\overline{p}\right]\left[2\beta-2\tau_1-\left(1+\alpha\right)d+\overline{p}\right]}$$

$$K_d = \frac{\left[\beta(\tau_1+\alpha d)-(\tau_1+\alpha d)^2\right]\left[\tau_2 - q\left[(1-\alpha)d-\overline{p}\right]\right]}{K\left[(1-\alpha)d-\overline{p}\right]\left[2\beta-2\tau_1-\left(1+\alpha\right)d+\overline{p}\right]}$$

Now, it only remains to select the filter time constant. A suitable choice of τ_F, is τ_F=τ_Z. With, this selection, the PD-1F controller settings in the case of large overshoot processes are obtained as suggested by the relations

$$K_I = \frac{\tau_2 - q\left[(1-\alpha)d\right]}{K\left[(1-\alpha)d\right]\left[2\beta-2\tau_1-\left(1+\alpha\right)d\right]} \tag{26}$$

$$K_P = \frac{\beta\left[\tau_2 - q\left[(1-\alpha)d\right]\right]}{K\left[(1-\alpha)d\right]\left[2\beta-2\tau_1-\left(1+\alpha\right)d\right]} \tag{27}$$

$$K_d = \frac{\left[\beta(\tau_1+\alpha d)-(\tau_1+\alpha d)^2\right]\left[\tau_2 - q\left[(1-\alpha)d\right]\right]}{K\left[(1-\alpha)d\right]\left[2\beta-2\tau_1-\left(1+\alpha\right)d\right]} \tag{28}$$

4 SIMULATION RESULTS

For assessment of the effectiveness of the proposed tuning methods and in order to provide a comparison with existing tuning methods, a series of simulation examples are carried out for different dead-time processes.

4.1 Inverse Response Processes with Two Stable Poles

Consider the typical inverse response process with K=1, τ_1=1, τ_2=1, d=0.8, p=-0.5, q=1. Applying the proposed method with α=0.6 and ξ_{des}= 0.8225, yields β=2.15. The PD-1F controller settings are then obtained as K_I=0.4221, K_P=0.9076 and K_d=0.4186. The settings of the series form PID controller with filtered derivative, tuned according the method proposed by Chien *et al* (2003), are K_C=0.3367, τ_I=1, τ_D=1, while the low-pass filter parameter takes the value a=0.1 and the inverse of the cyclic frequency of the desired critically damped closed-loop system takes the value τ_{cl}= 0.8348. The settings of the conventional PID controller that is tuned according to the method reported in Chen *et al* (2006), are K_C=0.71, τ_I=2, τ_D=0.5. Figure 3 illustrates the comparison of the servo-responses as well as of the regulatory control responses obtained by the proposed method and by the methods reported in Chien *et al* (2003) and Chen *et al* (2006), in the case of nominal process parameters. In case of set-point tracking, the proposed method provides a slightly more sluggish response as compared to the abovementioned PID tuning methods, while the initial jump obtained by our method is smaller. In the case of regulatory control, our method gives a better response in terms of maximum error, while the settling time is comparable to that obtained by the methods in Chien *et al* (2003) and Chen *et al* (2006).

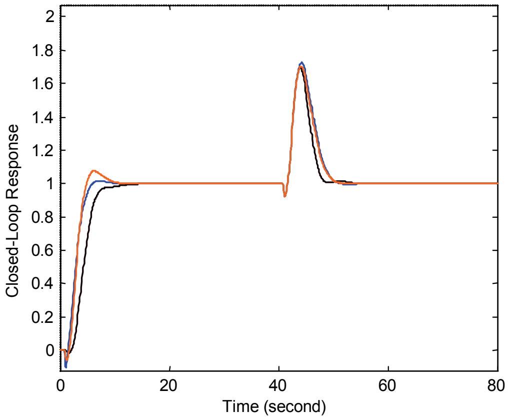


Figure 3: Servo-responses and regulatory control responses for the system $G_P(s)$=(-0.5s+1)exp(-0.8s)/(s+1), in case of nominal process parameters. Black line: Proposed Method. Orange line: Method in Chien *et al* (2003). Blue line: Method in Chen *et al* (2006).

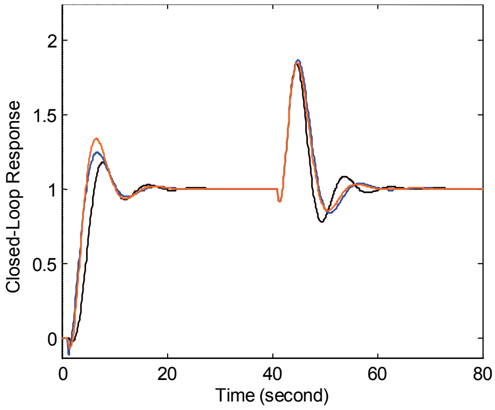


Figure 4: Servo-responses and regulatory control responses for the system $G_P(s)$=(-0.5s+1)exp(-0.8s)/(s+1), in case of a +20% mismatch in all process parameters. Other legend as in Figure 3.

A comparison in terms of the ISE criterion, in the case of regulatory control, gives the values 1.2002 for the proposed method, while for the methods in Chien *et al* (2003) and Chen *et al* (2006), we obtain ISE=1.5782 and ISE=1.4425, respectively. The respective IAE values for the methods under comparison are obtained as 2.443, 3.058 and 2.8941. Figure 4 shows the comparisons of the servo-responses and of the regulatory control responses in the case where a simultaneous +20% uncertainty in all process parameters is assumed. The responses obtained by the proposed method are better in terms of overshoot, maximum error and initial jump, while the settling time is similar to that of the responses obtained by the PID controllers tuned according to the methods by Chien *et al* (2003) and Chen *et al* (2006). The ISE values, in case of regulatory control, are 1.9514, for the proposed method, 2.3765 for the method of Chien *et al* (2003) and 2.1673, for the method of Chen *et al* (2006). The respective IAE values are 3.8221, 4.2492 and 3.9183.

As already mentioned, for a pre-specified value of adjustable parameter α, parameter β is directly related to the damping ration ζ of the second order approximation (10) of the closed-loop system. In

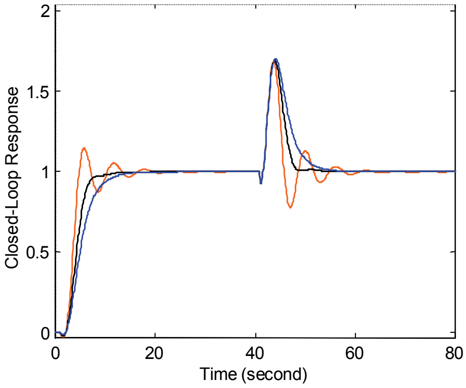


Figure 5: Servo-responses and regulatory control responses for the system $G_P(s)$=(-0.5s+1)exp(-0.8s)/(s+1), in case of nominal process parameters, for α=0.6 and for three values of β. Orange line: β=2.05; Black line: β=2.15;. Blue line: β=2.25.

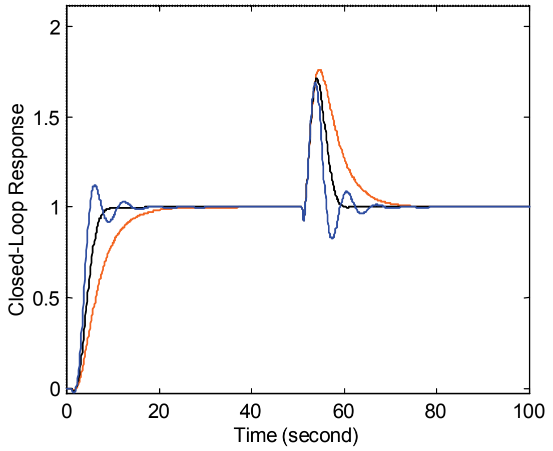


Figure 6: Servo-responses and regulatory control responses for the system $G_P(s)=(-0.5s+1)\exp(-0.8s)/(s+1)$, in case of nominal process parameters, for $\beta=2$ and for three values of α. Orange line: $\alpha=0.45$; Black line: $\alpha=0.5$;. Blue line: $\alpha=0.55$.

particular, as shown in Figure 5, β increases when ζ is increased. This of course results to a more conservative PD-1F controller. Therefore, a greater value of β, renders the closed-loop system more robust. Parameter α has an inverse effect on the closed-loop system robustness: For a pre-specified value of the parameter β, an increase of the parameter α, leads to a less robust but faster closed-loop system, as illustrated in Figure 6.

4.2 Control of a Continuous Stirred Tank Reactor

Let us consider the transfer function model of a CSTR reported in Padma Sree and Chidambaram (2004), and having the form

$$G_P(s)=\frac{-2.07(0.1507s+1)}{2.85s^2+2.31s-1}\exp(-0.3s)$$
$$=\frac{-2.07(0.1507s+1)}{(0.8905s+1)(3.2005s-1)}\exp(-0.3s)$$

The process has one dominant unstable pole and one stable pole, at s=0.3125 and s= -1.123, respectively, as well as a stable zero -6.6357. Here, K=-2.07, τ_1=0.8905, τ_2= 3.2005, d=0.3, p=0.1507, q=-1. Application of the proposed method with α=-0.5, β=2.5, yields the PD-1F controller settings K_P=-4.3862, K_I=-1.7545, K_d=-2.2859. The settings of the set-point weighted PID controller tuned according to the method reported in Padma Sree and Chidambaram (2004) are, K_C=-0.7205, τ_I=39.7228, τ_D=0.1494, while the tuning parameter used in the above mentioned paper, as well as the set-point weight b, take the values 0.15 and 0.3275, respectively. Figure 7 illustrates the servo-responses obtained by the two controllers. Figure 8 shows the comparison of the regulatory control responses for a negative unit step load change. Obviously, the PD-1F controller tuned according to the proposed method provides a considerably better performance, particularly in the case of regulatory control, where the response obtained by the controller tuned according to the method in Padma Sree and Chidambaram (2004) is practically unacceptable.

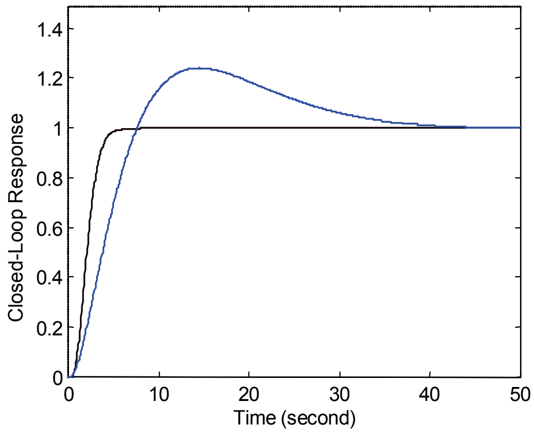


Figure 7: Closed-loop servo-responses of the CSTR model. Black line: Proposed method. Blue line: Set-point weighted PID controller tuned according to the method proposed by Padma Sree and Chidambaram (2004).

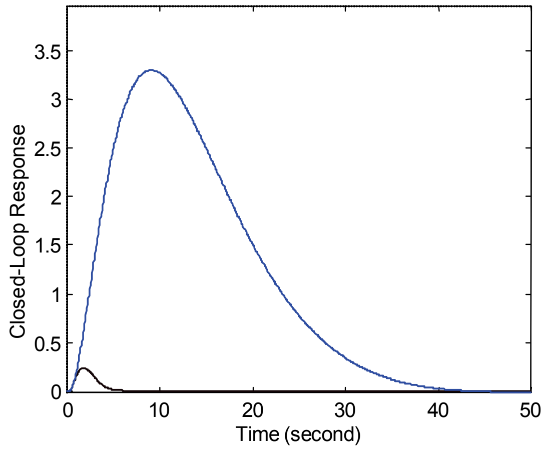


Figure 8: Regulatory control responses of the CSTR. Other legend as in Figure 7.

4.3 Second Order Unstable Process with a Positive Zero

Consider the process with K=1, τ_1=2.07, τ_2=5, d=0.939, p=-1, q=-1. The process has a stable pole, an unstable pole and a strong non-minimum phase zero. To the authors' best knowledge, controller design for second order processes with one or two righ-half-plane poles and a right-half-plane zero has not yet been addressed in the literature. Application of the proposed method, with α=0.3 and β=25, yields the PD-1F controller settings K_I=0.0920, K_P=2.3012, K_d=4.9027. The process model is next approximated as $G_P(s) = \exp(-1.939s)/\left[(2.07s+1)(5s-1)\right]$, i.e. the negative numerator time constant has been approximated as a time delay term of the form exp(-s). This is reasonable since an inverse response has a deteriorating effect on control similar to that of a time delay. We next apply the method reported in Lee *et al* (2000), in order to design a PID controller with first order set-point filter for the given process, on the basis of the approximated model. Application of the method reported in Lee *et al* (2000), with the IMC parameter λ=6.25, yields the PID controller settings K_C=1.9570, τ_I=34.9614 and τ_D=2.4889. Figure 9 illustrates the comparison of the servo-responses and the regulatory control responses for a unit step set-point change at t=0 sec and an inverse unit step load change at t=75 sec. It is seen that the proposed method results in an improved load disturbance response as compared to the method in Lee *et al* (2000), while the set-point responses are similar, with comparable settling times.

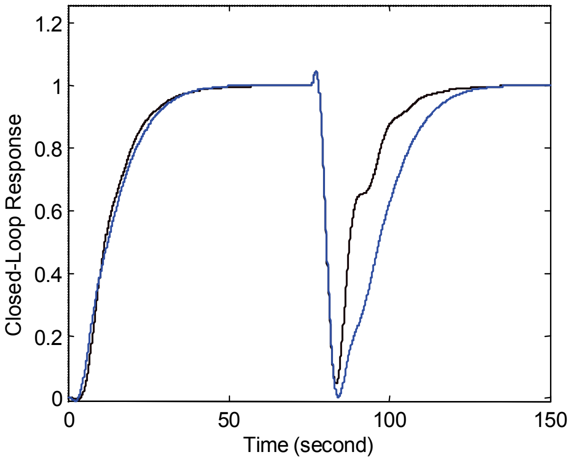


Figure 9: Servo-responses and regulatory control responses for the system G(s)=(-s+1)exp(-0.939s) / [(2.07s+1)(5s-1)]. Black line: Proposed method; Blue line: Method in Lee *et al* (2000).

4.4 Stable Second Order Unstable Process with a Positive Zero

Consider the process model of the form (2), with K=1, τ_1=2, τ_2=1, d=1, p=0.3, q=1. Application of the proposed method with α=0.4, β=3, yields K_I=1.0358, K_P=3.1073, K_d=1.6340. The settings of the series form PID controller with filtered derivative, tuned according the method proposed by Chien *et al* (2003) are K_C=1.0355, τ_I=2, τ_D=1, while the low-pass filter parameter takes the value τ_F=0.3 and the inverse of the cyclic frequency of the desired critically damped closed-loop system takes the value τ_{cl}= $d/\sqrt{2}$ = 0.5457. Figure 10 illustrates the comparison of the servo-responses as well as of the regulatory control responses obtained by the proposed method and by the method reported in Chien *et al* (2003). In the regulatory control case our method gives a considerably better response, whereas, although our method provides a smooth response, the method in Chien *et al* (2003) is better in the case of set-point tracking.

Let us now consider the case of a large overshoot process with K=1, τ_1=2, τ_2=1, d=1.2, p=5, q=1. Evaluating relations (17)-(19), while assuming α=0.2, β=3, yields the PD-1F controller settings K_I=0.2208, K_P=0.6624, K_d= 0.3759. Application of the above controller yields an unacceptable oscillatory response, as shown in Figure 11. Let us try, another design by evaluating relations (17)-(19) in the case where we select a=0.6, β=3. This yields K_I=0.1951, K_P=0.5853, K_d= 0.1486, i.e. a more conservative controller. The obtained servo-response

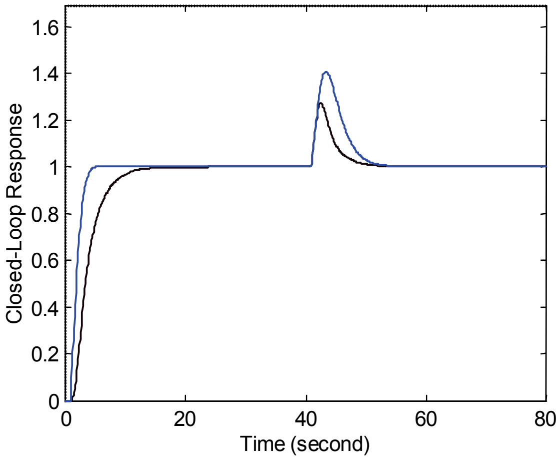


Figure 10: Servo-responses and regulatory control responses for the system $G_P(s)$=(0.3s+1)exp(-0.8s) /(2s+1)(s+1). Black line: Proposed method. Blue line: Method in Chien *et al* (2003).

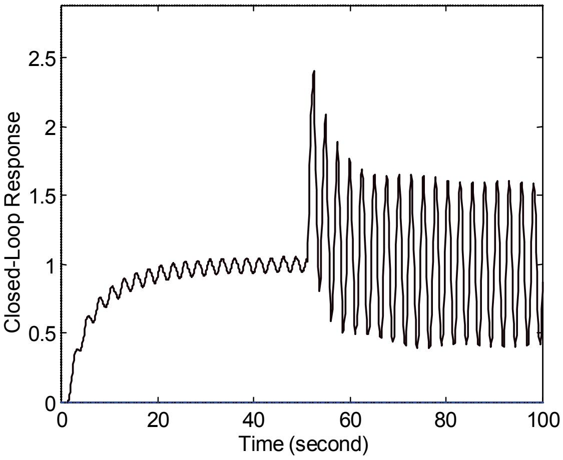


Figure 11: Closed-loop servo-response and regulatory control response of the system G(s)=(5s+1)exp(-1.2s) /(2s+1)(s+1), in the case of he PD-1F controller with parameters K_I=0.2208, K_P=0.6624, K_d= 0.3759.

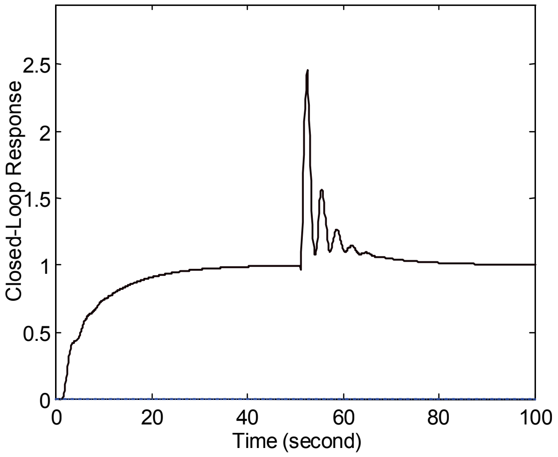


Figure 12: Closed-loop servo-response and regulatory control response of the system G(s)=(5s+1)exp(-1.2s) /(2s+1)(s+1), in the case of he PD-1F controller with parameters K_I=0.1951, K_P=0.5853, K_d= 0.1486.

and regulatory control responses are given in Figure 12. In the later case, the servo-response is quite smooth while the regulatory control response is less oscillatory. However, the robustness of the closed-loop system is marginal, and a small parameter mismatch can readily lead to instability.

Let us now consider filtering the output of the PD-1F controller that is designed for the case where α=0.2, β=3, with settings K_I=0.2208, K_P=0.6624, K_d= 0.3759, by a filter of the form (24), where τ_F=5. Moreover, let us design a PD-1F controller with filtered output as suggested by relations (26)-(28), with α=-0.2, β=2, τ_F=5. In this case the controller settings are K_I=0.3183, K_P=0.6366, K_d= 0.1344. Figure 13 shows the obtained servo-responses and regulatory control responses for both designs, together with the respective responses obtained by a series PID controller with filtered derivative, designed according the method reported in Chien *et al* (2003). It is seen that, in the regulatory control case our method gives a considerably better response, whereas, although our methods provide smooth responses, the method in Chien *et al* (2003) is better in the case of set-point tracking. A comparison in terms of ISE in the case of regulatory control gives the ISE values 1.8326 and 1.5892, for the proposed methods and 4.5754 for the method in Chien *et al* (2003). The respective IAE values are 4.5287, 3.7058 and 4.9453.

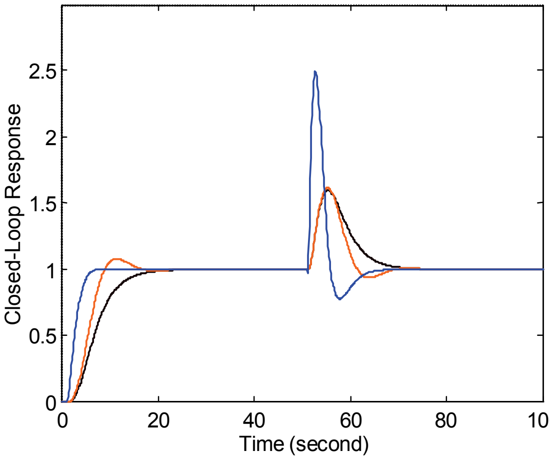


Figure 13: Closed-loop servo-response and regulatory control responses of the system G(s)=(5s+1)exp(-1.2s) /(2s+1)(s+1). Black line: PD-1F controller with filtered output tuned according to relations (17)-(19); Orange line: PD-1F controller with filtered output tuned according to relations (26)-(28); Blue line: Series form PID controller with filtered derivative tuned according to the method in Chien *et al* (2003).

5 CONCLUSIONS

A new direct synthesis method of tuning the PDF controller for stable or unstable dead-time processes with a negative or a positive zero has been presented. The proposed tuning method ensures smooth closed-loop response to set-point changes, fast regulatory control and sufficient robustness against parametric uncertainty. Numerical simulation examples verify the advantages of the proposed method over known PID controller tuning methods for the classes of dead-time processes under study. Extension of the proposed tuning method in the case of frequency domain specifications of the closed-loop system in terms of gain and phase margins is currently under investigation.

REFERENCES

Chang, D.M., Yu, C.C., Chien, I.-L., 1997. Identification and control of an overshoot lead-lag plant. *J. Chin. Inst. Chem. Eng*., 28, 79-89.

Chen, P.-Y., Tang, Y.-C., Zhang, Q.-Z., Zhang, W.-D., 2005. A new design method of PID controller for inverse response processes with dead time, *Proc. 2005 IEEE Conference on Industrial Technology (ICIT 2005),* Hong Kong, China, December 14-17, 2005, 1036-1039.

Chen, P.-Y., Zhang, W.-D., Zhu, L.-Y., 2006. Design and tuning method of PID controller for a class of inverse response processes. *Proc. 2006 American Control Conference,* Minneapolis, Minnesota, U.S.A., June 14-16, 2006, 274-279

Chien, I.-L., Chung, Y.-C., Chen B.-S., Chuang, C.-Y., 2003. Simple PID controller tuning method for processes with inverse response plus dead time or large overshoot response plus dead time. *Ind. Engg. Chem. Res.,* 42, 4461-4477.

Lee, Y.-H., Lee, J.-S., Park, S.-W., 2000. PID controller tuning for integrating and unstable processes with time delay. *Chem. Engg. Sci.,* 55, 3481-3493.

Luyben, W.L., 2000. Tuning Proportional-Integral Controllers for processes with both inverse response and deadtime. *Ind. Eng. Chem. Res.,* 39, 973-976.

Padma Sree, R., Chidambaram, M., 2004. Simple method of calculating set point weighting parameter for unstable systems with a zero. *Comp. Chemical Engg.,* 28, 2433-2437.

Phelan, R.M., *Automatic Control Systems,* New York, Cornell University Press, 1978.

Scali, C., Rachid, A., 1998. Analytical design of Proportional-Integral-Derivative controllers for inverse response processes. *Ind. Eng. Chem. Res.,* 37, 1372-1379.

Waller, K.V.T., Nygardas, C.G., 1975. On inverse response in process control. *Ind. Eng. Chem. Fundam.,* 14, 221-223.

Zhang, W., Xu, X., Sun, Y., 2000. Quantitative performance design for inverse-response-processes. *Ind. Eng. Chem. Res.,* 39, 2056-2061.

A DISCRETE EVENT SIMULATION MODEL FOR THE EGRESS DYNAMICS FROM BUILDINGS

Paolo Lino, Bruno Maione
Dept. of Electrical and Electronics Engineering, Technical University of Bari, Bari, Italy
lino@deemail.poliba.it, maione@poliba.it

Guido Maione
Dept. of Environmental Engineering and Sustainable Development, Technical University of Bari, Taranto, Italy
gmaione@poliba.it

Keywords: Evacuation dynamics, Queuing networks, Discrete event systems, SimEvents©.

Abstract: Safe egress of people from closed buildings is a critical issue, in which modern control methodologies and information and communication technologies play a crucial role. Current research trends suggest us to profitably use wireless networks of distributed sensors and actuators. Then, a large amount of feedback from the real scenario is needed to determine control outputs. In this paper, we use a discrete event system approach to define a simulation model of a complex real scenario. The egress of students and academic staff from a lecture area in the School of Engineering in Bari was simulated, to validate the modeling approach in predicting the evacuation process. Performance indices (flows of individuals in spaces and at critical points, number of evacuated people, time to complete egress) were measured in standard conditions when no emergency or panic phenomena occurred. The results show that the model properly represents real phenomena like blocking, congestion or overcrowding, and faster-is-slower effect. Then, the same approach could be efficient to predict flows in emergency conditions, when specific control actions are taken for speeding-up egress safely.

1 INTRODUCTION

Recently, safe egress of people from large buildings in standard or emergency conditions has received considerable attention. In particular, after the the 9/11 Twin Towers terrorist attack in New York City the evacuation of complex and/or high buildings has been a focus of attention. All the world over, safety has been based on prescriptive design regulations concerning building characteristics (distances, number of exits, exit widths, etc.) which allow the occupants to evacuate the structure within a pre-defined acceptable amount of time. Hence evacuation procedures assuring acceptable, building safety standards have been a major concern for engineers. Consequently, the crowd management has been based on the assessment of the people handling capability of building spaces before using them. However, the shortcomings of this strategy is that it fails to take into account how people actually behave during the egress. Today there is a tendency to control the behavior of occupants also before, during and after the evacuation process. However, the egress control for influencing the behavior implies a research effort both in mathematical modeling and information and communication technologies (ICT).

The mathematical or simulation models can be very useful in: describing the crowd dynamics during evacuation by means of system parameters (*e.g.* crowd distribution and speed); studying critical phenomena (blocking and congestion); measuring performance indices (number of evacuated individuals, time required, speed, etc.); designing and optimizing buildings and escape routes; comparing alternative control strategies. A good control strategy to route individuals should predict and dynamically adapt itself to the different emergency conditions, the different and random distributions and behaviors of individuals (type and time of reaction to alarms, decisions taken, etc.), the random events (interruption of escape routes, doors or exits blocked, overcrowding close to emergency exits, etc.). Then, suitable control actions are based on feedback from the environment.

Sensing and communication technologies are used to measure variables which can indicate emergency and/or panic, and, at the same time, to communicate actions for safely escaping from the risky environment. Such communications can be directed to all people by using distributed actuators (monitors, flashing lights, automatically opening doors, acoustic signals and alarms, etc.), or to specific expert human agents, devoted to help and direct groups of people to a safe exit, by using Personal Digital Assistants (PDAs) or palmtop computers.

Recently, our research group started up a scientific project to profitably use wireless sensor networks and ICT for managing evacuation from buildings during emergencies. The main goal is reducing egress times in a safe way. After a literature review, a model suitable to develop supervisory control policies and a test-bed are currently under investigation. The model of the crowd dynamics defines the feedback information and the control actions. In particular, the time required to manage an emergency condition is $T = T_1 + T_2 + T_3$, given by the time to feel and recognize emergency (T_1), the time to elaborate sensed information (T_2), the time to route the crowd in a safe condition (T_3). Control should minimize T_3.

Scientific literature reports flow-based models using graphs or similar tools, cellular automata, agent-based systems in which agents represent individuals, activity-based models including sociological and behavioral aspects (Schreckenberg and Sharma, 2003; Santos and Aguirre, 2004; Kuligowski and Peacock, 2005; Waldau et al., 2007). Flow-based models are mostly based on the *carrying capacity, i.e.* they predict the evacuation dynamics by considering the topology of the building or physical location in which the emergency occurs, and the evacuation policies (Schreckenberg and Sharma, 2003). Other models consider also the *human response*, *i.e.* the psychological or sociological factors, and individual reactions (Galea et al., 1996; Klüpfel et al., 2000; Schadschneider et al., 2008). The two modeling approaches differ for a macroscopic or microscopic point of view, respectively.

Macroscopic models are usually employed to statically plan escape routes, for achieving the 'quickest flow' or the 'maximum flow', and they are not adapted by the feedback from the real scenario. Neither microscopic models can be adapted in real time, because a dynamic optimization of escape routes and flows would require too much computational resources and time. Moreover, a detailed microscopic simulation environment could require information that can't be acquired during emergency. Basically, macroscopic models do not consider individual characteristics and behaviors, but they synthesize a common emerging behavior. On the contrary, microscopic models consider each individual as an autonomous decision making entity, moving and behaving according to both personal and general criteria.

Then, we built a model useful to control evacuation in real time, on the basis of the information needed and control outputs. Important state feedback is about: distribution and number of individuals in the evacuated areas; measured flows in critical points, and congestion or overcrowding of specific areas or points that reduce flow; binary condition (crossable/not crossable) of routes, doors, exits, transit points, which can be affected by fire, smoke, structural problems, etc.. Typical control outputs can be associated to: flashing lights showing the best direction to a safe exit; acoustic signals; automatic opening of doors to a safe exit, and automatic closing of doors to dangerous or critical areas; instructions and orders given by expert operators.

Asynchronous events occurring in emergency conditions, and the discrete nature of controlled variables and signals from actuators, justify using a discrete event system (Cassandras and Lafortune, 1999) to model, analyze, and control the evacuation of people. Typical events are sudden variation of available paths, blocking of doors, elevators out of service, automatic closing/opening of doors, etc..

In particular, queuing networks (Kleinrock, 1975) easily describe precedence relations, parallelism, synchronization, modularity, and other properties. More specifically, they can be used to statistically represent the decisions and actions affecting the evacuated crowd behavior. To this aim, a probabilistic approach may take into account several decision parameters, which depend on the current system state and are related to sociological and psychological factors. The human decision is based on elaboration of perceived signals and information, not simply on a causal *stimulus-reaction* relation. For example, consider when individuals interact and form groups, or try to rescue relatives going in opposite direction to the crowd, or the influence of leaders, expert agents, firemen, and so on. This approach simplifies the control system design, and, at the same time, considers an individual perspective to a certain extent. Moreover, escape routes can be easily recognized, and minimum time/shortest length paths can be identified.

State dependent queues in the proposed model make it difficult to find a closed form solution for performance analysis. Thus, a simulation model has been implemented in MATLAB/Simulink$^{\copyright}$ environment, by means of the discrete event simulation tool *SimEvents*. Here, we report some results on a case-study used to test our approach, based on queuing networks and discrete event systems theory.

Section 2 briefly introduces the model and the assumptions made. Section 3 describes the developed simulation model. Section 4 gives the performance measured in the simulated case-study. Section 5 draws the conclusions.

2 THEORETICAL MODEL OF THE EGRESS DYNAMICS

Here, we summarize the assumptions made to build a discrete event system model of the crowd dynamics in standard or emergency conditions. We represented the phenomenon as a queueing network system, composed by different queues, each one describing the behavior of individuals in a zone of the evacuated environment. A zone could be a room, a corridor, a stairway, an exit or an entrance, a door, but also a floor or level of a building. Then, the approach can be used to model and simulate complex networked buildings and environments, by integrating and connecting different queues in a single representation.

In this framework, we described the behavior of people as an elementary queue with parameters determined by physical human peculiarities, according to the Kendall notation (Kleinrock, 1975). The queue service rate is interpreted as the time necessary to cross rooms, corridors, stairs, and depends on the *free walking speed*, *i.e.* the speed an individual may reach in an open space. This speed is function of age, sex, physical conditions and abilities, external pressure to hurry, dawdling, baggage carried, gradient of walking area (Fruin, 1971; Tregenza, 1976). An average value $v_0 = 1.34\ m/s$ and a standard deviation of 0.26 for a normal distribution are commonly accepted (Weidmann, 1993). But actual walking speed is nonlinearly affected by density ρ of individuals. Experimental studies showed that the *average impeded speed* v decreases as the number of persons P per unit area increases (Fruin, 1971; Tregenza, 1976): ρ has almost no influence up to 0.27 P/m^2, and motion is stopped when $\rho_{max} = 5\,P/m^2$ (Tregenza, 1976), which is taken as maximum space capacity. A linear relation can be assumed between v and ρ, if $\rho \in [0.3, 2]$. Here, we assume the motion of individuals in rooms and corridors as described in (Weidmann, 1993), according to the following formula:

$$v(\rho) = v_0 \left[1 - e^{-\gamma\left(\frac{1}{\rho} - \frac{1}{\rho_{max}}\right)}\right], \qquad (1)$$

where $\gamma = 1.913$ is a fit parameter.

For motion on stairways, we consider the free 'horizontal' speed, *i.e.* the horizontal component of the speed vector, as normally distributed. The average is function of the previously cited parameters and of the stair geometry (angle and riser height). Short and long stairways can be distinguished (Fruin, 1971; Kretz et al., 2008): the first exhibit higher speeds when walking down-up, the latter when going up-down. In this paper, we assume short stairways traveled in both directions (average free up-down speed 0.780 m/s, average free down-up speed 0.830 m/s), and long stairways only down-up (average free speed 0.423 m/s). These v_0 values (Kretz et al., 2008) are used for the impeded actual speed in (1).

Moreover, interactions between individuals increase with ρ, especially in bottlenecks (Helbing et al., 2000). Frictions occur when people wish to move faster than the currently achieved speed, a typical panic behavior. Then, arch-like clusters form and grow at doors, exits, or other critical points, if desired walking speed v_d exceeds the critical free walking speed (Helbing et al., 2000; Parisi and Dorso, 2007). The consequence is a *faster-is-slower* effect which delays the egress. Then, two different outflow regimes exist depending on v_d: the first is when outflow depends linearly on v_d (the faster individuals want to move, the faster they evacuate); the second is when outflow decreases with v_d, due to interactions.

Queues with null queueing space and a certain server capacity are used to represent rooms, corridors, stairways, doors, exits, entrances and gateways. Each queue can accommodate as many people as the capacity of the modeled space (Jain and Smith, 1997). If a unit space has a capacity of 5 P/m^2, an area of length L and width W has a capacity $C = 5 \cdot L \cdot W$. The service time is normally distributed, with an average value given by $L/v(\rho)$. Differences between the modeled spaces are obtained by specifying a different v_0 for each type of space. Arrivals to queues are exponentially distributed, as it is commonly assumed and also observed. Summing up, we obtain state dependent $M/G/C/C$ queues.

In particular, doors, exits, entrances, and gateways are modeled by queues with a server capacity equal to the width W of the passage (more precisely the maximum number of individuals that can flow through). If the way is filled at its capacity, then the queue of the antecedent space is blocked.

The queue service rate is determined by taking into account the faster-is-slower effect, as described in the following. First of all, it is supposed that the desired walking speed of individuals crossing a bottleneck varies as proposed by (Helbing et al., 2000):

$$v_d(t) = [1 - p(t)]\,v_d(0) + p(t)\,v_d^{max}, \qquad (2)$$

where $v_d(0)$ is the initial desired speed, v_d^{max} is the maximum desired speed, and $p(t)$ specifies the crowd impatience (Helbing et al., 2000), with:

$$p(t) = 1 - \frac{\overline{v}(t)}{v_d(0)}, \tag{3}$$

being $\overline{v}(t)$ the average speed of individuals in the crowd. Then, we assume that the queue desired service rate $\mu_d(t)$ and the average service rate $\overline{\mu}(t)$ relate to the desired and average speeds according to $\mu_d(t) = W v_d(t)$ and $\overline{\mu}(t) = W\overline{v}(t)$, respectively. Finally, the actual service rate μ is normally distributed with an average value given by (Wang et al., 2008):

$$E[\mu \mid \mu_d] = \begin{cases} \mu_d & if \quad \mu_d \le \mu_c \\ 1 - e^{\frac{\alpha}{\mu_d - \mu_c}} & if \quad \mu_d > \mu_c \end{cases} \tag{4}$$

where $E[\mu \mid \mu_d]$ is the expected value of the service rate μ, μ_c is the flow capacity of the passage, and α is a negative constant. To sum up, firstly μ_d is computed and compared to μ_c, then $E[\mu \mid \mu_d]$ is used to generate μ.

3 THE SIMULATION MODEL

The proposed model represents the main aspects of the evacuation process, and can be exploited to carry out performance analysis in terms of egress times, number of evacuees per time unit, length of queues, existence of bottlenecks and congestion. Unfortunately, a closed form solution giving the steady state probabilities of the network cannot be easily found, as service times depend on the system state. Moreover, the real time management of evacuation can take advantage from the knowledge of the transient dynamics, which cannot be analytically determined. Thus, a queueing network simulation model providing a tool suitable for implementing and validating evacuation strategies is developed in the MATLAB/Simulink© environment. In particular, we exploit the discrete event system toolbox *SimEvents*. Just like other software tools like Arena, Extend, Witness, etc., it allows the representation of complex discrete-event systems by a network of queues. Moreover, the integration with MATLAB and Simulink simplifies the modeling process of hybrid dynamical systems, which include continuous-time, discrete-time and discrete-event subcomponents, such as sensor networks and distributed control systems.

Figure 1 depicts the block scheme of the queue which models wide areas, like rooms and corridors. We assume the flow in one direction. The main elements of the scheme are a *FIFO queue* representing the queueing space and a *N-server*, consisting of a number of servers matching the available capacity.

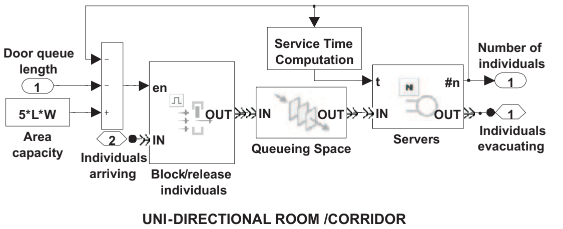


Figure 1: *SimEvents* implementation of rooms, corridors, and stairways.

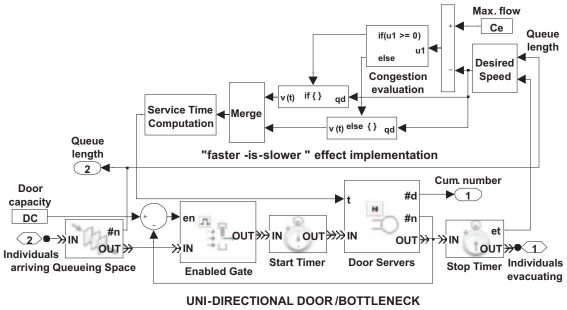


Figure 2: *SimEvents* implementation of bottlenecks.

The function *Service Time Computation* computes the service time depending on the area congestion. It consists of two functions: the first derives the current speed from (1) by considering the number of people crossing the area; the second computes the service time as the path length divided by the speed. The *Block/release* element prevents individuals to enter area, if the maximum capacity $5 \cdot L \cdot W$ has been reached.

For stairways we use the same scheme in Figure 1: free walking speeds specified in Section 2 are used in (1) to compute the current speed in congestion conditions. More precisely, the individual space occupancy is suitably increased for upward motion, because people oscillate sideways when rising stairways, which reduces the available space.

The block scheme implementing bottlenecks like doors is represented in Figure 2, and it suitably models the faster-is-slower effect. The model is composed of a *FIFO queue*, whose space will be defined in the next subsection, and a *N-server* with as many servers as the individuals that can cross the bottleneck at the same time. The service time is determined by (4), (2) and (3), provided that an estimate of the average service rate $\overline{\mu}(t)$ is available. If ΔT is the time interval taken by the last individual to cross the door, as measured between blocks *Start Timer* and *Stop Timer*, its reciprocal $\mu = 1/\Delta T$ represents the current service rate. Thus, since the number n of individuals waiting to be served has a zero service rate, the overall aver-

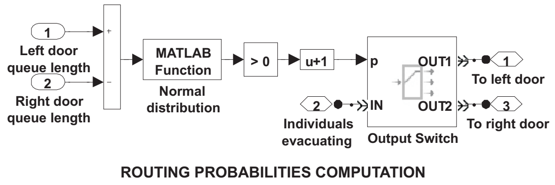


Figure 3: Transition from corridors/rooms to multiple doors.

age service rate can be computed as:

$$\overline{\mu} = \frac{\sum \mu_i}{(n+1)} = \frac{1}{(n+1)\cdot\Delta T} \qquad (5)$$

Then, the *Desired speed* block calculates $p(t)$ according to (3) and $v_d(t)$ according to (2). If the resulting value overcomes the door maximum flow capacity, a congestion occurs. Finally, the *Service time computation* block outputs a service time obtained from a normal distribution with a mean equal to the reciprocal of the service rate.

When rooms/corridors and doors share the same queueing space, we must guarantee that the number of individuals in the system does not overcome the overall capacity. Then, the door queueing space capacity is set equal to the room/corridor capacity. So, after connecting the elementary sub-models, the number of individuals waiting in front of the door is used to reduce the number of available servers in the room/corridor. As an example, individuals arriving at the end of a corridor enter the door queue and wait for a free server. At the same time, they reduce the corridor available space, but do not affect the walking speed of individuals crossing the corridor. To implement this condition, the signal *Door queue length* representing the number of individuals in the door queue is fed back (see Figure 1).

For rooms/corridors with more than one door an *Output switch* block connects the area to exits. The routing probability is set for each possible direction (Figure 3). We assume that probability to choose each door is inversely affected by its crowding condition. To introduce a sufficient level of uncertainty in the choice, we generate a number from a normal distribution having the length of each queue as its average. Then, the selection comes from comparing the results.

4 SIMULATION RESULTS

As a case-study, we consider the area of large lecture rooms at Technical University of Bari, *i.e.* 5 lecture rooms and a Great Hall, all connected to a main corridor, which has an entrance/exit point 2.73 *m* wide and a maximum flow of 3 persons at a time (Figure 4).

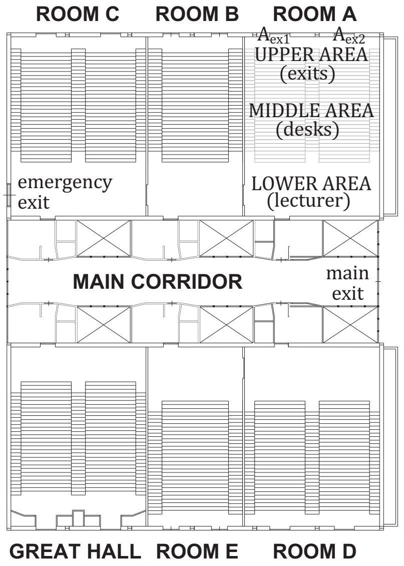


Figure 4: The case-study.

The Hall is 294 m^2 large, with a maximum capacity of 270 persons. Three rooms (A, C, D) are 294 m^2 large, with a maximum capacity of 270 persons. Two smaller rooms (B, E) are 207 m^2 large, with a maximum capacity of 180 persons. Sitting desks in the Great Hall and in A, C, D are vertically distributed from a lower to an upper level, an internal corridor separates desks in two columns and two more external corridors are available. Rooms B, E have only one column of desks and two external corridors. All rooms have one single access/exit point at the lower level (1.6 *m* wide, maximum flow of 2 persons at a time), used by academic staff, and two access/exit doors at the upper level (2.3 *m* wide, maximum flow of 2 persons at a time), used by students. The lower level doors link rooms to the main corridor, which is 235 m^2 large. Each room communicates with its adjacent room(s), except for the Great Hall: the three communication doors are 2.3 *m* wide. Room C has also a further emergency exit (see Figure 4).

To sum up, there are 14 points of exit: one from the main corridor, 12 from the upper level doors, one from room C. Then, the main and natural flow of students during evacuation is through the upper doors, otherwise through the corridor, especially the ones sitting in the first lines of desks. The teaching staff can use the room lower exit doors, the corridor and then its exit. People in room C can use the added emergency exit, which is an opportunity also for people in the Great Hall (e.g. if the exits from the Great hall are blocked or unavailable). Each room is divided into 3 main areas, representing the lecturer (lower), the desks (middle), and the exit (upper) areas, respec-

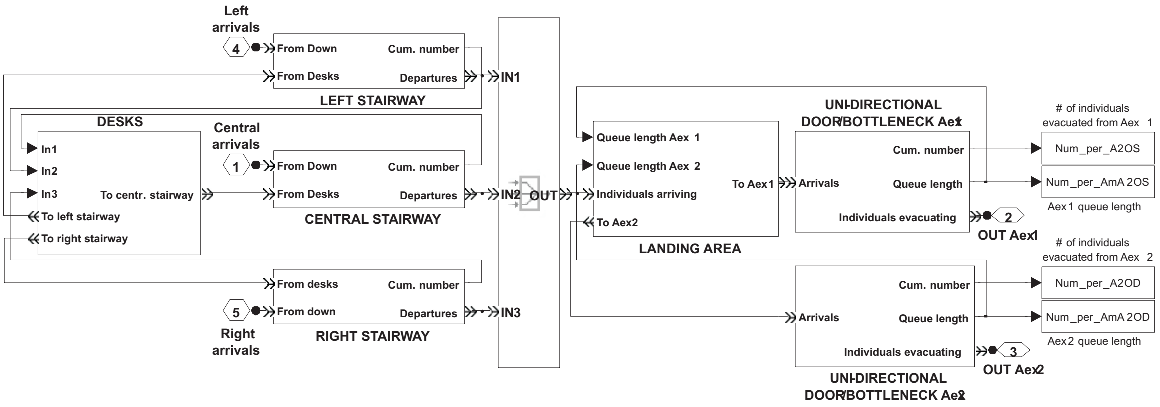


Figure 5: SimEvents model for upper and middle areas in room A.

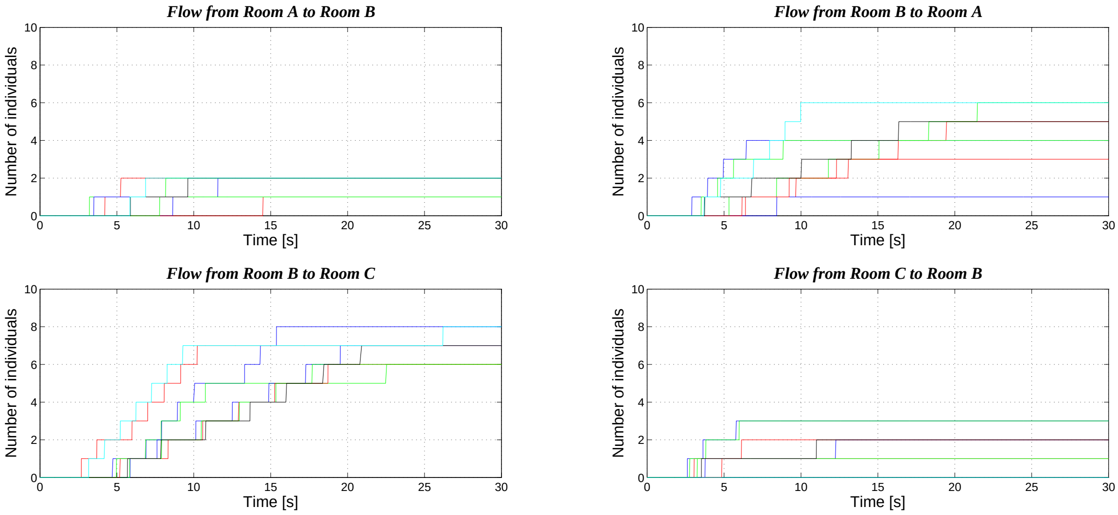


Figure 6: Cumulative number of individuals crossing doors connecting Rooms A, B and C.

tively. The exit area consist of a landing space receiving individuals from stairways and includes two exits. Then, 3 queues are associated to the lower area, 2-3 queues to the middle area, depending on the number of staircases, and 3 to the last area, *i.e.* two for exits and one for the landing space.

The *SimEvents* block scheme for the upper and middle areas of room A is in Figure 5.

An extensive simulation analysis has been executed to predict the evacuation dynamics. Only relevant results are presented. Without loss of generality, we assume evacuation in normal circumstances, *i.e.* panic or environmental conditions do not affect the behavior. Representation of evacuation under panic conditions simply needs a tuning of model parameters, which is under investigation.

We suppose that egress starts at the end of a lecture session, so that all the rooms are evacuated simultaneously. As initial condition, an average population of 150 individuals occupies each room, mainly distributed in the desks area, while the lower and upper areas are sparsely populated. The Great Hall and the main corridor are initially empty. We assume that individuals occupying the desk and upper areas evacuate from exits in the same room, while those in lower areas evacuate from the starting room or toward an adjacent room or the corridor.

All results refer to 5 different simulation runs. Figure 6 represent flows trough doors connecting rooms A and B, and B and C, respectively. The flow is composed of individuals initially occupying the lower areas.

It is evident that only few of them try to evacuate from larger rooms A and C toward room B, because routing probabilities depend on the current crowding. In fact, the initial crowd density is larger in the smaller rooms. Finally, the cumulative number of individuals go from a room to another one changes at each

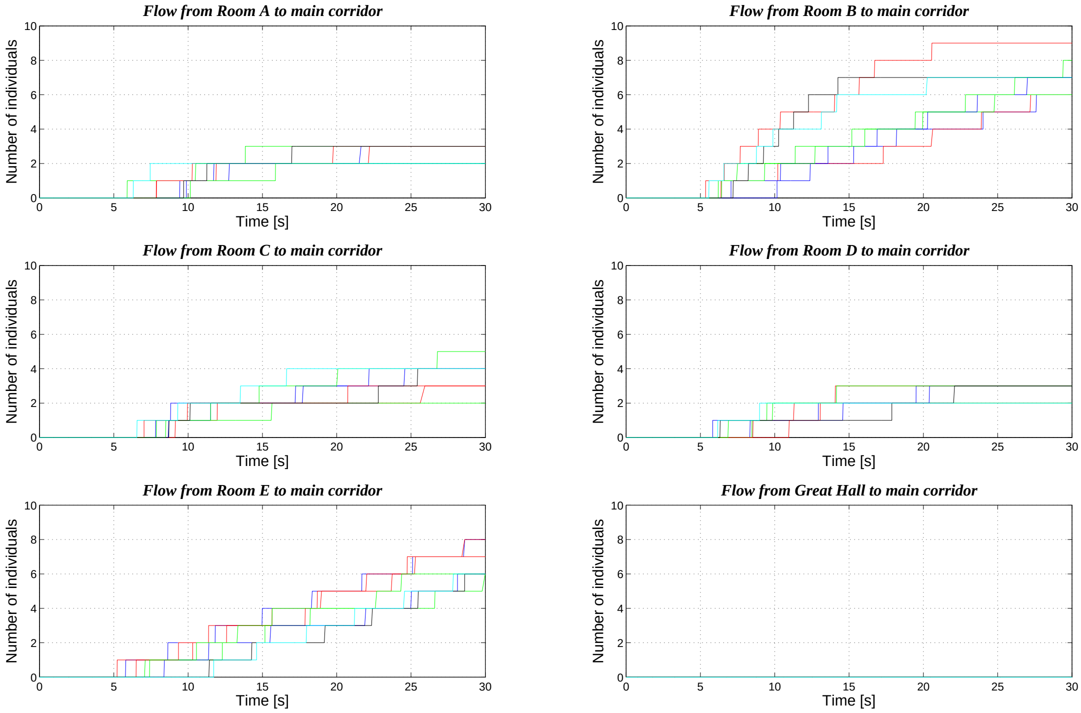
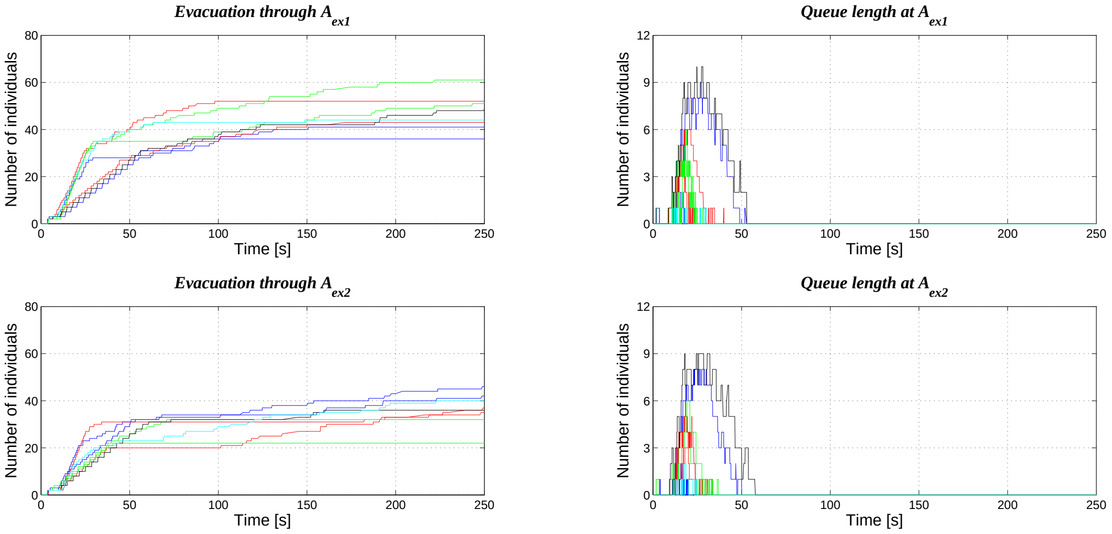


Figure 7: Cumulative number of individuals flowing from Rooms and Great Hall to main corridor.



(a)

(b)

Figure 8: Evacuation from Room A: (a) cumulative number of individuals evacuating from A_{ex1} and A_{ex2}; (b) queues length at A_{ex1} and A_{ex2}.

run, due to the randomness of the transitions. Similar flows are obtained for other rooms.

Flows of individuals choosing the main corridor are shown in Figure 7.

Cumulative numbers of individuals increase almost linearly in all cases, being the doors capacities sufficient to handle the traffic. Just after 30 *s* all individuals have abandoned lower areas.

Finally, Figure 8 depicts evacuation from Room A through the two upper exits.

The flow is mainly composed of individuals leaving the desk area. Figure 8(a) shows that most of people leaves after a delay of about 10 *s*, which is nearly the time necessary to cover half of the stair length. During the initial transient, the individuals reaching exits can immediately evacuate with a minimum service time, as doors are initially free. Conversely, slopes of curves in Figure 8(a) reduce with overcrowding of queues in the upper area, which delay individuals. Figure 8(b) shows that individuals reaching the upper area direct themselves almost uniformly towards A_{ex1} and A_{ex2}, as the choice is affected by the doors crowding. After 60 *s*, the queues of upper area and exits are nearly empty, so that arriving individuals are promptly served. The overall evacuation takes 120-160 *s* on average.

5 CONCLUSIONS

In this paper, a simulation model describing the evacuation dynamics from buildings has been presented, considering the queueing network theory as a modeling tool. The model is suitable for implementing and testing control strategies for managing emergency situations. Results from a simulation model implemented in the Matlab/Simulink$^{©}$ environment, by using the discrete events simulation toolbox *SimEvents*, have shown the feasibility of the approach. Without loss of generality, simulation represents evacuation dynamics in ordinary conditions. Parameters tuning for panic situations is under investigation. A further validation is under development by comparing preliminary results with those obtained using commercial tools.

ACKNOWLEDGEMENTS

This work is part of the Research Project ”*Infrastrutture di telecomunicazione e reti wireless di sensori nella gestione di situazioni di emergenza*” funded by Apulia Regional Government, Italy.

REFERENCES

Cassandras, C. and Lafortune, S. (1999). *Introduction to discrete event systems*. Kluwer Academic Publishers, Norwell, USA.

Fruin, J. (1971). *Pedestrian - Planning and Design*. Metropolitan Association of Urban Designers and Environmental Planners, New York, USA.

Galea, E., Owen, M., and Lawrence, P. (1996). he exodus evacuation model applied to building evacuation scenarios. *Fire Engineers Journal*, 6:27–30.

Helbing, D., Farkas, I., and Vicsek, T. (2000). Simulating dynamical features of escape panic. *Nature*, 407:487–490.

Jain, R. and Smith, J. (1997). Modeling vehicular traffic flow using m/g/c/c state dependent queueing models. *Transportation Science*, 31(4):324–336.

Kleinrock, L. (1975). *Queuing Systems, volume I: Theory*. John Wiley & Sons, New York, USA.

Klüpfel, H., Meyer-Konig, T., and Schreckenberg, M. (2000). Microscopic simulation of evacuation processes on passenger ships. In *Proceedings of the 4th International Conference on Cellular Automata for Research and Industry*, pages 63–71, Karlsruhe, Germany.

Kretz, T., Grünebohm, A., Kessel, A., Klüpfel, H., Meyer-König, T., and Schreckenberg, M. (2008). Upstairs walking speed distributions on a long stairway. *Safety Science*, 46:72–78.

Kuligowski, E. and Peacock, R. (2005). A review of building evacuation models. Technical Note 1471, NIST, USA.

Parisi, D. and Dorso, C. (2007). Morphological and dynamical aspects of the room evacuation process. *Physica A*, 385:343–355.

Santos, G. and Aguirre, B. (2004). A critical review of emergency evacuation simulation models. In *NIST Workshop on Building Occupant Movement during Fire Emergencies*, pages 25–50, USA. NIST Press.

Schadschneider, A., Klingsch, W., Klüpfel, H., Kretz, T., Rogsch, C., and Seyfried, A. (2008). *Encyclopedia of Complexity and System Science*, chapter Evacuation Dynamics: Empirical Results, Modeling and Applications. Springer, Berlin, Germany. To appear.

Schreckenberg, M. and Sharma, S., editors (2003). *Pedestrian and Evacuation Dynamics*. Springer-Verlag, Berlin, Germany.

Tregenza, P. (1976). *The Design of Interior Circulation*. Crosby Lockwood Staples, London, UK.

Waldau, N., Gatterman, P., Knoflacher, H., and Schreckenberg, M., editors (2007). *Pedestrian and Evacuation Dynamics 2005*. Springer-Verlag, Berlin, Germany.

Wang, P., Luh, P., Chang, S., and Sun, J. (2008). Modeling and optimization of crowd guidance for building emergency evacuation. In *4th IEEE Conference on Automation Science and Engineering*, pages 328–334, Key Bridge Marriot, Washington DC, USA.

Weidmann, U. (1993). Transporttechnik der fussgänger - transporttechnische eigenschaften des fussgngerverkehrs. London, uk, Institut füer Verkehrsplanung, Transporttechnik, Strassen - und Eisenbahnbau IVT an der ETH Zürich. In German.

MODELING, SIMULATION AND FEEDBACK LINEARIZATION CONTROL OF NONLINEAR SURFACE VESSELS

Mehmet Haklidir, Deniz Aldogan, Isa Tasdelen and Semuel Franko
TUBITAK Marmara Research Centre, Information Technologies Institute, 41470, Gebze-Kocaeli, Turkey
mehmet.haklidir@bte.mam.gov.tr, deniz.aldogan@bte.mam.gov.tr
isa.tasdelen@bte.mam.gov.tr, semuel.franko@bte.mam.gov.tr

Keywords: Surface Vessels, Nonlinear Analysis and Control, Feedback Linearization.

Abstract: Realistic models and robust control are vital to reach a sufficient fidelity in military simulation projects including surface vessels. In this study, a nonlinear model including sea-state modelling is obtained and feedback linearization control is implemented in this model. To control the system, nonlinear analysis techniques are used. The model is integrated into a commercial framework based CGF application within a high-fidelity military training simulation.The simulation results are presented at the end of the study.

1 INTRODUCTION

The aim of this study is to observe the dynamic behaviors of the surface vessels under the effect of hydrodynamic force-moments and environmental conditions such as waves, current, wind, season that pertaining to the tactical environment.

The analysis and control of nonlinear motion model of surface vessels are obtained by using following techniques:

- Linearization by Taylor Series
- Phase Plane Analysis
  - o Course Keeping
  - o Zig Zag Maneuver
- Lyapunov Stability Theorem
- Feedback Linearization

Ship dynamics model and disturbance model are introduced in Section 2; the phase plane analysis and lyapunov stability therom in Section 3 and 4, the proposed controller is discussed in Section 5; simulation results are presented in Section 6.

2 THE SURFACE PLATFORM MOTION MODULE

2.1 Coordinate System and Vector Notation

The motion of surface vessels has 6 degrees of freedom. The description and notation of each degree of freedom has been shown on Table 1.

Table 1: DoF Description and Notation.

DOF	Description	Axis	Forces and moments	Linear and angular veloc.	Positions and Euler angles
1	Surge	x	X	u	x
2	Sway	y	Y	v	y
3	Heave	z	Z	w	z
4	Roll	x	K	p	ϕ
5	Pitch	y	M	q	θ
6	Yaw	z	N	r	ψ

SNAME's (1950) notation is used in this study. The first three parameters and time derivatives that are shown on Table define the position and the motion of the platform in x-, y-, z- axes. Last three parameters define the orientation and rotary motion of the platform. After analyzing 6 degrees of freedom motion of surface vessel, it is observed that 2 axis systems are needed to perform the motion. Therefore, North – East- Down (NED), is the local geodetic coordinate system fixed to the Earth, and Body Fixed, is fixed to the hull of ship, coordinate frames are used. Motion axis system X_0 Y_0 Z_0 has been fixed to the platform and called as Body Fixed axes system. The point O, which is the origin of this axes system, is always selected as the ship's centre of gravity. (Figure 1)

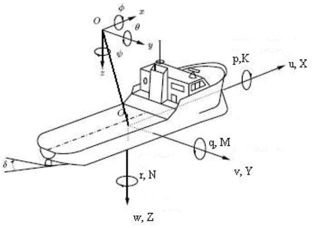


Figure 1: Coordinate System.

2.2 Surface Platform Motion Equations

Fossen (1991), by inspiring Craig's (1989) robot model, contrary to classical representation, modeled 6 degrees of freedom motion of the surface vessel vectorially.

$$\dot{\boldsymbol{\eta}} = J(\boldsymbol{\eta})\boldsymbol{\upsilon}$$

$$M\dot{\boldsymbol{\upsilon}} + C(\boldsymbol{\upsilon})\boldsymbol{\upsilon} + D(\boldsymbol{\upsilon})\boldsymbol{\upsilon} + \mathbf{g}(\boldsymbol{\eta}) = \boldsymbol{\tau} + \mathrm{g}_0 + \mathrm{w}$$

Above; M is the moment of inertia including added mass, C(υ) is Coriolis matrix, D(υ) is damping matrix, g(η) is gravitational force vector and τ is the vector showing the force and moments of the propulsion system that causes motion. This representation will be used in this study.

2.2.1 Motion Equations

Representing the motion equations in the Cartesian system of coordinates (body-fixed reference frame) and defining x_G, y_G and z_G as the position of the ship's CG, the well known motion equations of a rigid body are giving by the following (Fossen, 1991):

Surge:

$$X = m[\dot{u} + qw - rv + x_G(q^2 + r^2) + y_G(pq - \dot{r}) + z_G(rp + \dot{q})]$$

Sway:

$$Y = m[\dot{v} + ru - pw - y_G(r^2 + p^2) + z_G(qr - \dot{p}) + x_G(qp + \dot{r})]$$

Heave:

$$Z = m[\dot{w} + pv - qu - z_G(p^2 + q^2) + x_G(rp - \dot{q}) + y_G(rq + \dot{p})]$$

Roll :

$$K = I_X\dot{p} + (I_Z - I_y)qr + m[y_G(\dot{w} + pv - qu) - z_G(\dot{u} + ru - pw)]$$

Pitch:

$$M = I_y\dot{q} + (I_x - I_Z)rp + m[z_G(\dot{u} + qw - rv) - x_G(\dot{w} + pv - qu)]$$

Yaw:

$$N = I_Z\dot{r} + (I_y - I_x)pq + m[x_G(\dot{v} + ru - pw) - y_G(\dot{u} + qw - rv)]$$

2.2.2 Simplifying Assumptions

Simplifying assumptions used in this study are following:

- The rotational velocity and acceleration about the y-axis are zero (q, = 0).
- The translational velocity and acceleration in the z direction are zero. (w, = 0).
- The vertical heave and pitch motions are decoupled from the horizontal plane motions.
- The vertical centre of gravity, (VCG), is on the centerline and symmetrical (yG=0)

2.2.3 Simplified Motion Equations

Applying simplifying assumptions to the general motion equations, the following simplified equations of motion are obtained

Surge: $$X = m[\dot{u} - rv - x_G r^2 + z_G rp] \quad (1)$$

Sway: $$Y = m[\dot{v} + ru - z_G\dot{p} + x_G\dot{r}] \quad (2)$$

Roll: $$K = I_X\dot{p} - mz_G(\dot{u} + ru)] \quad (3)$$

Yaw: $$N = I_Z\dot{r} + mx_G(\dot{v} + ru) \quad (4)$$

2.2.4 Force and Moments Acting on Surface Vessel

Basically force and moments acting on surface vessel can be divided to four as; hydrodynamics force and moments, external (environmental) loads, control surface forces (rudder, fin..) and propulsion (propeller) forces. Force and moments can be expressed according to axis system;

Surge: $X = X_H + X_R + X_E + T$

Sway: $Y = Y_H + Y_R + Y_E$

Roll: $K = K_H + K_R + K_E$

Yaw: $N = N_H + N_R + N_E$

Description of indices is; H, Hydrodynamic force and moments originating from fluid-structure interaction, R , forces that affects control surface are, E, environmental external loads (Wave, current, wind), T, propulsion force.

Hydrodynamic Forces and Moments

Integration of the water pressure along the wetted area of the surface vessel causes hydrodynamic force and moments within the platform. These force and moments can be defined, with the velocity and acceleration terms as a nonlinear axes system, by using Abkowitz method.

Most important step on developing maneuver model is expanding force and moment terms in Taylor's series. This way, nonlinear terms act as independent variables and form a polynomial equation. The function and its derivatives have to be continuous and finite in the region of values of the variables to use the Taylor's expansion. Certainty of the model alters depending on where the expansion is finished.

Force and moments, which were obtained by expanding Taylor series until third power, are under mentioned (Abkowitz, 1969; Sicuro, 2003)

$$X_{hid} = X_{\dot{u}}(\dot{u}) + X_{vr}vr + X_{uu}v^2 \quad (5)$$

$$\begin{aligned} Y_{hid} &= Y_{\dot{v}}\dot{v} + Y_{\dot{r}}\dot{r} + Y_{\dot{p}}\dot{p} + Y_{\phi|uv|}\phi|uv| + Y_{\phi|ur|}\phi|ur| \\ &+ Y_{\phi|uu|}\phi|uu| + Y_{|u|v}|u|v + Y_{ur}ur + Y_{v|v|}v|v| \quad (6) \\ &+ Y_{v|r|}v|r| + Y_{r|v|}r|v| \end{aligned}$$

$$\begin{aligned} K_{hid} &= K_{\dot{v}}\dot{v} + K_{\dot{p}}\dot{p} + K_{|u|v}|u|v + K_{ur}ur + K_{v|v|}v|v| \\ &+K_{v|r|}v|r| + K_{r|v|}r|v| + K_{|uv|\phi}|uv|\phi \\ &+K_{|ur|\phi}|ur|\phi + K_{uu\phi}uu\phi + K_{|u|p}|u|p \\ &+K_{|p|p}|p|p + K_p p + K_{\phi\phi\phi}\phi\phi\phi - \Delta G_z(\phi) \end{aligned} \tag{7}$$

Obtaining Hydrodynamic Derivatives

In order to obtain hydrodynamic derivatives three basic methods can be used.

- ➢ By means of basin test using the realistic model
- ➢ By using CFD (Computational Fluid Dynamics) software
- ➢ By using empirical formulae

In this study third method was used. Hydrodynamic derivatives have been used by the empirical formulae of the source Inoue et al. (1981). To have an opinion about validity and fidelity of the empirical formulae, parameters of a merchant ship that was chosen from literature was used. By using these parameters and related formulae hydrodynamic derivatives were calculated and compared with the equivalent in the literature.(Table 2)

Table 2: Comparing the hydrodynamic derivatives obtained from the model data and empirical formulae.

	Model (Son and Nomoto, 1982)	Emprical Formulas (Inoue et. al, 1981)
Yv	-0.0116	-0.0118
Nv	-0.0038	-0.0041
Yr	0.0024	0.0022
Nr	-0.0022	-0.0020
Yrvv	0.0214	0.0216
Nrvv	-0.0424	-0.0427
Yrrv	-0.0405	-0.0426
Nrrv	0.0016	0.0017

The Environmental Disturbances

The environmental disturbances acting on the surface vessels can be grouped into two main categories; the wave model, the current and wind models.

The Wave Model

When real data regarding the complicated seas lacks, idealized mathematical spectrum functions are generally used for marine calculations. One of the easiest and commonly used of these calculations is the Pierson – Moskowitz spectrum where a wave spectrum formula is provided for winds blowing over an infinite area and at a constant speed for over a sea of full state. In this study this spectrum is used while a wave model is created.(Berteaux, 1976)

This spectrum is expressed as follows due to the wave frequency and wind speed.

$$\mathrm{S}_{\xi} = \frac{0.0081\mathrm{g}^2}{\omega^5}\exp\left[-0.74\left(\frac{\mathrm{g}}{\mathrm{V}\omega}\right)^4\right] \tag{8}$$

where, ω : Wave Frequency [rad/sec], V: Wind Speed (at 19,5 m above sea) [m/s]

The Current and Wind Models

Typically wind models only treat the force and moments that are directly related to surge, sway and yaw motions. In this study, the wind model is obtained by using Isherwood Method.(Isherwood 1972)

Wind forces and moments acting on a surface platform are usually defined in terms of relative wind speed V_R (knots) and relative angle γ_R (deg). The wind forces for surge and sway and the wind moment for yaw as is shown.

$$X_{wr} = \frac{1}{2}C_X(\gamma_R)\rho_w V_R^2 A_T \tag{9}$$

$$Y_{wr} = \frac{1}{2}C_Y(\gamma_R)\rho_w V_R^2 A_L \tag{10}$$

$$N_{wr} = \frac{1}{2}C_N(\gamma_R)\rho_w V_R^2 A_L L \tag{11}$$

where C_X, C_Y and C_N are the force and moment coefficients, ρw is the density of the air, AT and AL are the transverse and lateral projected areas and L is the overall length of the ship. (Isherwood, 1972). The equations of current forces and moments are similar with wind forces and moments.

2.3 Nonlinear Equations of Motion

When the simplified 4 degrees of freedom motion model, which was obtained in previous section, was associated with hydrodynamic forces and environmental external loads a nonlinear maneuver model can be obtained. To behave like independent variables and become coefficients of a polynomial motion equation, hydrodynamic derivatives are derived by another software that makes use of ship geometry. As well, terms of ship motion equations are normalized relative to the ship velocity. (Fossen, 1991)

$$\begin{aligned} X' &= X'(u') + (1-t)T'(J) + X'_{vr}v'r' + X'_{vv}v'^2 \\ &+X'_{rr}r'^2 + X'_{\phi\phi}\phi'^2 + c_{RX}F'_N\sin\delta' \end{aligned} \tag{12}$$

$$\begin{aligned} Y' &= Y'_v v' + Y'_r r' + Y'_p p' + Y'_\phi \phi' + Y'_{vvv}v'^3 + Y'_{rrr}r'^3 \\ &+Y'_{vvr}v'^2 r' + Y'_{vrr}v'r'^2 + Y'_{vv\phi}v'^2\phi' + Y'_{v\phi\phi}v'\phi'^2 \\ &+Y'_{rr\phi}r'^2\phi' + Y'_{r\phi\phi}r'\phi'^2 + (1+a_H)F'_N\cos\delta' \end{aligned} \tag{13}$$

$$\begin{aligned} K' &= K'_v v' + K'_r r' + K'_p p' + K'_\phi \phi' + K'_{vvv}v'^3 + K'_{rrr}r'^3 \\ &+K'_{vvr}v'^2 r' + K'_{vrr}v'r'^2 + K'_{vv\phi}v'^2\phi' + K'_{v\phi\phi}v'\phi'^2 \\ &+K'_{rr\phi}r'^2\phi' + K'_{r\phi\phi}r'\phi'^2 - (1+a_H)z'_R F'_N\cos\delta' \end{aligned} \tag{14}$$

$$\begin{aligned} N' = {} & N'_v v' + N'_r r' + N'_p p' + N'_\phi \phi' + N'_{vvv} v'^3 + N'_{rrr} r'^3 \\ & + N'_{vvr} v'^2 r' + N'_{vrr} v' r'^2 + N'_{vv\phi} v'^2 \phi' + N'_{v\phi\phi} v' \phi'^2 \\ & + N'_{rr\phi} r'^2 \phi' + N'_{r\phi\phi} r' \phi'^2 + (x'_R + a_H x'_H) F'_N \cos\delta' \end{aligned} \qquad (15)$$

3 PHASE PLANE ANALYSIS

3.1 Phase Portrait of Course Keeping

Phase portraits of surface platform are shown. Yaw angle (psi) versus its derivative yaw rate (r) in Figure 2 and Roll angle versus roll rate in Figure 3 are used to obtain the phase portraits. If the real part of the eigenvalues is positive, then x(t) and x(t) both diverge to infinity, and the singularity point is called an unstable focus.

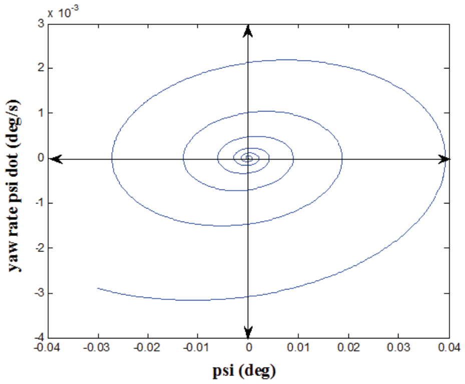


Figure 2: The Phase Portrait (Yaw vs Yaw rate).

The phase portrait in Figure 3 demonstrates that the unstable free motion of the surface platform.

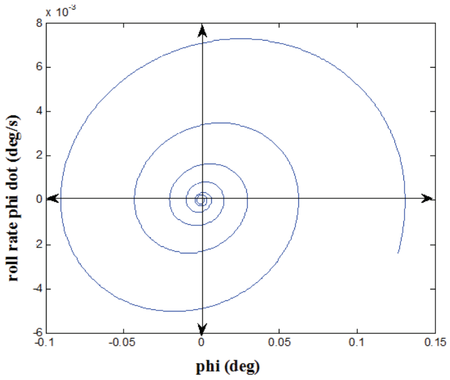


Figure 3: The Phase Portrait(Roll angle vs Roll rate).

3.2 Phase Portrait of Zig Zag Maneuver

It is intended that the surface platform makes zig-zag maneuvers of 45° with a velocity of 8 m/s with 20° rudder angle. For a zig-zag maneuver, when the angular acceleration plotted is against angular velocity it shows how non-linear ship response can be (Figure 4).

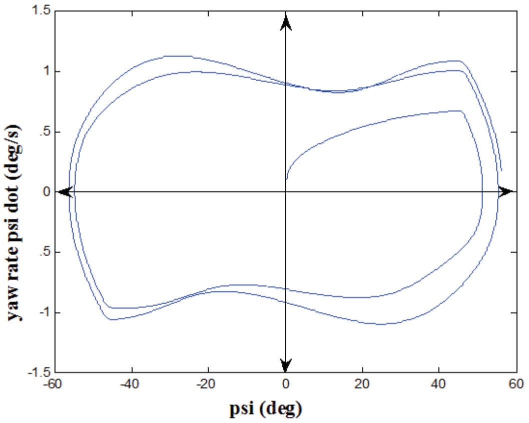


Figure 4: Phase Portrait of Zig Zag Maneuver.

4 LYAPUNOV STABILITY THEOREM FOR SURFACE PLATFORM DYNAMIC

A fully actuated surface platform can be described by

$$M\dot{\boldsymbol{\upsilon}} + C(\boldsymbol{\upsilon})\boldsymbol{\upsilon} + D(\boldsymbol{\upsilon})\boldsymbol{\upsilon} + \mathbf{g}(\boldsymbol{\eta}) = Bu = \boldsymbol{\tau}$$
$$\dot{\boldsymbol{\eta}} = J(\boldsymbol{\eta})\boldsymbol{\upsilon}$$

where J(η) is singular for θ = ±90 degrees (Euler angles), M= M^T>0 and D(ν) = D^T(ν) > 0. The position is controlled by

$$u = B^T (BB^T)^{-1} \left[g(\eta) - J^T(\eta) K_P \eta \right]$$

where Kp = K^Tp > 0. Let $V = \frac{1}{2}\left(\upsilon^T M \upsilon + \eta^T K_P \eta\right)$ be a Lyapunov function candidate for the closed-loop system (4.1), (4.2) and (4.3). We take the time derivative of the Lyapunov function candidate to obtain

$$\begin{aligned} \dot{V} & = \upsilon^T \left(M\dot{\upsilon} + J^T(\eta) K_P \eta \right) \\ & = \upsilon^T \left(Bu - C(\upsilon)\upsilon - D(\upsilon)\upsilon - g(\eta) + J^T(\eta) K_P \eta \right) \\ & = \upsilon^T \left(-C(\upsilon)\upsilon - D(\upsilon)\upsilon \right) \\ & = -\upsilon^T D(\upsilon)\upsilon \end{aligned}$$

which is ***negative semidefinite***. Asymptotic stability can then be established by applying LaSalle's invariance principle, but the equilibrium point (η, ν)=(0, 0) is only ***locally asymptotically stable*** since J(η) is singular for θ = ±90 degrees.

5 FEEDBACK LINEARIZATION

The basic idea with feedback linearization is to transform the nonlinear systems dynamics into a linear system (Freund (1973). Conventional control techniques like pole placement and linear quadratic optimal control theory can then be applied to the linear system. Feedback linearization allows us to design the controller directly based on a nonlinear dynamic model that better describes a ship maneuvering behavior. Consider Norrbin's nonlinear ship steering equations of motion in the form (Fossen 1992):

$$m\ddot{\psi} + d_1\dot{\psi} + d_3\dot{\psi}^3 = \delta \qquad (16)$$

here m = T/K, d_1 = n_1/K and d_3 = n_3/K. Taking the control law to be:

$$\delta = \hat{m}a_\psi + \hat{d}_1\dot{\psi} + \hat{d}_3\dot{\psi}^3 \qquad (17)$$

where the hat denotes the estimates of the parameters and a, can be interpreted as the commanded acceleration, yields:

$$m(\ddot{\psi} - a_\psi) = \tilde{m}a_\psi + \tilde{d}_1\dot{\psi} + \tilde{d}_3\dot{\psi}^3 \qquad (18)$$

Here $\tilde{m} = \hat{m} - m$, $\tilde{d}_1 = \hat{d}_1 - d_1$ and $\tilde{d}_3 = \hat{d}_3 - d_3$ are the parameter errors. Consequently, the error dynamics can be made globally asymptotically stable by proper choices of the commanded acceleration a_ψ. (Fossen 1992) In the case of no parametric uncertainties, equation (18) reduces to: $\ddot{\psi} = a_\psi$ which suggests that the commanded acceleration should be chosen as:

$$a_\psi = \ddot{\psi}_d - K_d\dot{\tilde{\psi}} - K_p\tilde{\psi} \qquad (19)$$

where ψ_d is the desired heading angle and $\tilde{\psi} = \psi - \psi_d$ is the heading error. This in turn yields the error dynamics:

$$\ddot{\tilde{\psi}} + K_d\dot{\tilde{\psi}} + K_p\tilde{\psi} = 0 \qquad (20)$$

The block diagram of the control system is shown in Figure 5.

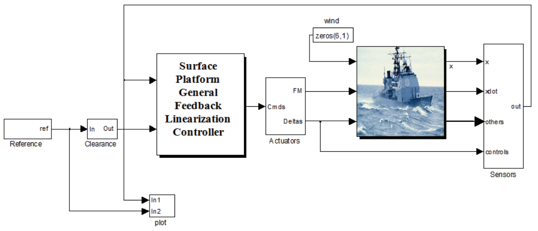


Figure 5: Block Diagram of System.

6 EXPERIMENTAL RESULTS

The crucial parameters of the surface platform chosen for the illustration have been displayed in Table 3.

Table 3: The main parameters of the surface platform.

Description		Value	Description	Value
Total Length		171 m	Volume	12000 m^3
Total Width		20.4 m	Height	12 m
Draft	Front	5.9 m	Block Coefficient	0.559
	Rear	5.7 m	Area of Rudder	28 m^2
	Middle	5.8 m	Beam/Length Ratio	1.8219

In the sample application, it is intended that the surface platform makes zig-zag maneuvers of 45° with a velocity of 8 m/s. The route information regarding this task is inputted by the VR-Forces graphical user interface (Figure 6).The results below have been produced after running the simulation for 800 seconds.

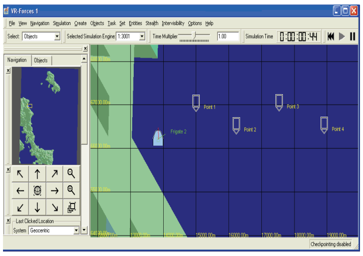


Figure 6: The route defined for the platform.

In this application, which is known as the zig zag test of Kempf in the literature (Kempf, 1932), the initial speed of the platform has been given as 0. The platform is ordered to move to the specified waypoints one by one by increasing its velocity up to 8 m/s. It takes the platform 96 seconds to reach to the first point. The first loop is accomplished in approximately 295 seconds. The results are acceptable for the motion behaviors that are supposed to be realized by a large platform and satisfactory in terms of simulation.

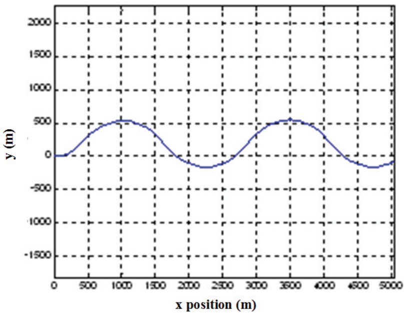


Figure 7: (a) Change of location.

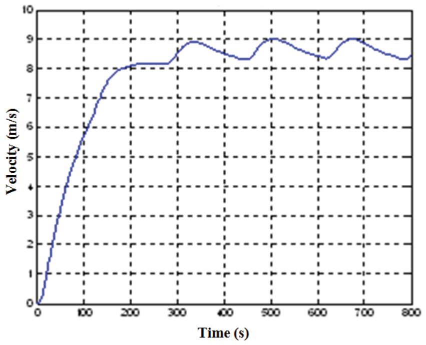


Figure 7: (b) Change of velocity.

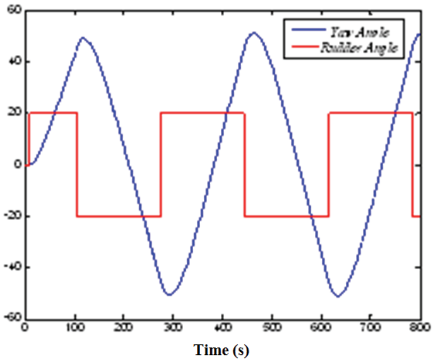


Figure 8: Changes in the yaw angle and rudder angle of the surface platform.

Controller performance can tried by some different route applications:

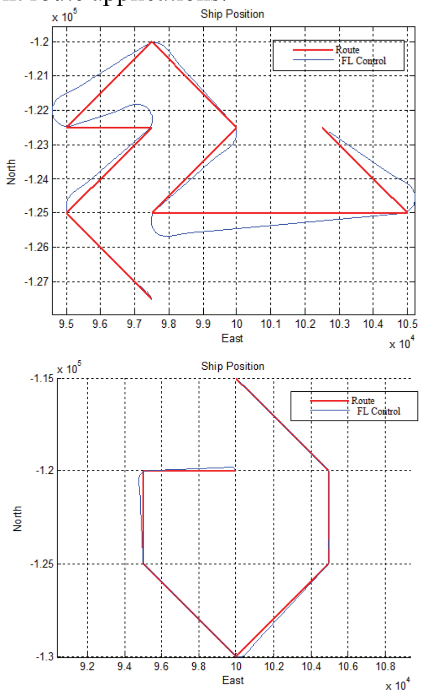


Figure 9: Controller performance in different routes.

7 CONCLUSIONS

In this study, feedback linearization control has been implemented in a nonlinear surface vessel model including sea-state modeling (wave, current, wind). The performance of the maneuver controller has been illustrated through a simulation study. The results are acceptable and satisfy for the needs of military simulation. Although we have designed our control to cover all influences, a more specified design can upgrade the performance in each different case. In the future work, the performance of the controller may be compared with an intelligent control technique.

REFERENCES

Abkowitz, M. A. , 1969. Stability and Motion Control of Ocean Vehicles, M.I.T. Press, Cambridge, Massachusetts.

Berteaux, H. O. , 1976. Buoy Engineering, Wiley and Sons, New York.

Fossen, T. I., 1991. Nonlinear Modeling and Control of Underwater Vehicles, Dr. Ing. thesis, Dept. of Engineering Cybernetics, The Norwegian Institute of Technology, Trondheim.

Fossen, T. I. and Paulsen, M. J., 1992. Adaptive Feedback Linearization Applied to Steering of Ships, Proceedings of the 1st IEEE Conference on Control Applications (CCA'92), Dayton, Ohio, September 13-16, 1992, pp. 1088-1093.

Freund, E., 1973. Decoupling and Pole Assignment in Nonlinear Systems. Electronics Letter, No.16.

Inoue, S., Hirano, M., Kijima, K., 1981. Hydrodynamic derivatives on ship manoeuvring; International Ship Building Progress, Vol. 28.

Isherwood, R. M. , 1972. Wind Resistance of Merchant Ships, RINA Trans., Vol. 115,pp. 327-338.

Kempf, G., 1932. Measurements of the Propulsive and Structural Characteristics of Ships, Transactions of SNAME, Vol. 40, pp. 42-57.

SNAME, 1950. The Society of Naval Architects and Marine Engineers. Nomenclature for treating the motion of submerged body through a fluid, Technical Research Bulletin No. 1-5

MODELING OF CONTINUOUS FERTILIZER GRANULATION-DRYING CIRCUIT FOR COMPUTER SIMULATION AND CONTROL PURPOSES

Gediminas Valiulis and Rimvydas Simutis
Department of Process Control, Kaunas University of Technology, Studentų St. 48, Kaunas, Lithuania
gvaliulis@gmail.com, rimvydas.simutis@ktu.lt

Keywords: Granulation, Modeling, Simulation, Control.

Abstract: The paper presents the model-based approach to process simulation and advanced control in the industrial granulation circuit of fertilizer production. Different knowledge sources, such as physical phenomena, statistical analysis of process parameters, expert information cover different cognition domains of the process. The mechanistic growth model developed is based on particle coating phenomena, mass and energy transfer. The model partially takes into account the main process parameters, features and the equipment used. Simulation has been executed to test the model performance. The model built can be used for the evaluation of plant control methods and staff training.

1 INTRODUCTION

Drum granulation is a commonly used process in a commercial fertilizer production. Many continuous granulation plants operate well below design capacity, suffering from high recycle rates and even periodic instabilities (Wang and Cameron, 2002). The main reasons are related to raw material properties, process equipment and control problems.

The process control still depends on the experience and skills of process operators, namely experts. Diagnostic systems show potential to apply systems engineering approaches to complex operational problems such that operators are well informed, are able to quickly diagnose abnormal conditions, test quickly possible solutions via detailed simulations and then proceed to apply corrective actions (Salmon et al., 2007). However, a number of interacting process variables (some of them are stochastic in nature) lead to a complex dynamic system that might be hard to predict and optimize just by intuition, especially for unskilled operators. Fortunately, it is possible to use granulation process simulations provided by PC for the investigation of such complex problems.

The aim of this paper is to propose the process simulator based on an extended modeling approach for continuous drum granulation-drying processes, focused on simulation and control. This approach involves the dynamic process model built from heterogeneous knowledge sources such as physical principles, empirical (measured) data and expert information.

The mechanistic part incorporates the understanding of physics and underlying mechanisms (e.g. mass and energy balances, growth kinetics).

The empirical part uses raw and/or filtered process sensors' data, their storage, retrieval and parameter identification techniques in addition to the mechanistic (white box) model.

The expert component involves the process experts' recommendations, which are of great value due to the lack of other knowledge mentioned above.

2 MAIN PROCESS DETAILS

Drum granulation is a particle size enlargement process often obtained by spraying a liquid binder or slurry onto fine particles as they are agitated in a rotary drum (Wang and Cameron, 2002). The particle circulation is achieved mechanically (by the action of the rotating drum and lifters). Granules are cycled many times through the spray zone and the liquid layer attached is pre-dried before the particle returns to the spray zone again (Figure 1).

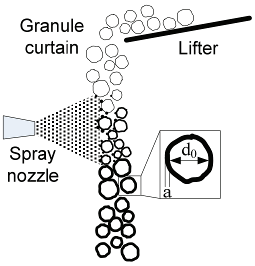


Figure 1: View of the coating phenomenon.

The desired mode of granule growth is layering (coating), resulting in very tight granule size distributions.

A commercial continuous granulation circuit for granulated diammonium phosphate fertilizer (formed by the reaction of phosphoric acid and ammonia) production consists of the following major parts: a pipe reactor, spray nozzle system, drum granulator-dryer, granule classifier (screens), crusher and nuclei feed system (Figure 2).

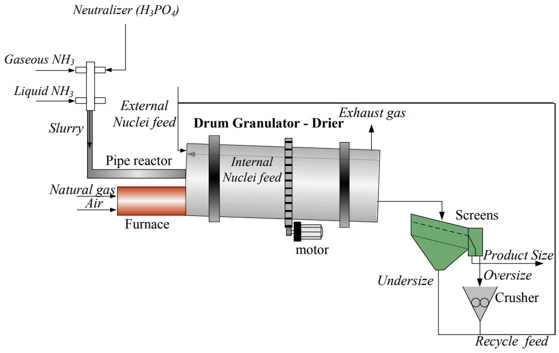


Figure 2: Typical drum granulation-drying circuit utilized in the diammonium phosphate (DAP) production industry.

A granulation drum is made of an inclined cylinder with simultaneous drying (there is no separate drying device). Drying is performed by the heat of burned natural gas and/or reaction heat of phosphoric acid and ammonia. Liquid DAP feed (slurry) is sprayed onto the tumbling bed of seeds via spraying nozzles. The drum is tilted lengthwise a few degrees to provide the flow of granules through the drum length. The backward screw sends a part of granules (internal nuclei) back to the spraying zone. Granules from the granulator-drier are transmitted to the classifier and split into three fractions: undersize, oversize, and marketable product size. The oversize fraction is crushed and sent back to the granulator together with undersize granules.

Fortunately, nowadays some important granule size distribution variables can be measured on-line using advanced particle size analysis systems. Detailed and more accurate information provides the producers of granulated materials with more data to improve product quality and to control production processes. Size Guide Number (SGN), related to the median of granule population, and Uniformity Index (UI), which shows the dispersion of population, can be evaluated. A part of important granule size distribution intervals can be also provided.

However, some process variables connected with material and equipment properties can not be evaluated and controlled directly. In such a situation the process model can provide information about important process states, such as recycle size flow rate and distribution, drum system jamming factor, granule moisture content, size evolution of single granule inside the granulator-dryer. This information can help to predict future process states and prevent abnormal situations, which can initiate process stoppage and loss of productivity.

3 MODELING

The model presented here is essentially based on fundamental conservation principles, with partial consideration of equipment properties and the stochastic nature of the process. For modeling purposes, it is necessary to divide the granulation circuit into several balance areas with the central component of the model – the drum granulator-drier. There are two main processes inside the granulator-drier: the growth of particles and moisture evaporation (drying).

Basic modeling assumptions are:

- granule shape is spherical;
- each granule in the granulation circuit is analyzed;
- stochastic nature of the process is estimated;
- preferred growth is by layering;
- granule agglomeration is an unacceptable mode of operation;
- growth rate is a function of initial granule size, slurry flow rate, temperature inside the granulator, granule position in the drum, number of particles in the granule bed;
- mechanical attrition of granules inside the granulator-drier is defined by attrition function;
- presumable nucleation (formation of new seeds) occurs during slurry spraying;
- external classification of granules into three fractions (undersize, marketable and oversize) is defined by classification function;

- external crushing of oversize granules is characterized by grinding function;
- residence and transportation delays in the plant are considered;
- internal and external seeds serve as nuclei for new granules.

3.1 White Box Modeling

There are two basic granule growth mechanisms that act independently or in combination (Findlay et al., 2005). A successive layering of binding material on an initial nucleus is termed layering, coating or ''onion-skin'' growth mechanism. Another mechanism is an agglomeration or coalescence process that occurs upon particle collision. Whereas growth by agglomeration mostly occurs when a binder is added, layered growth is the result of particle coating by the feed material, followed by solidification of the material on the particle surface (Degreve et al., 2006).

The granulation regime depends on some factors such as slurry viscosity and purity, N:P mole ratio, granule curtain density, temperature of slurry and seed to be coated, granule density, air temperature inside the granulator-dryer, etc. Some of these parameters can be observed and controlled, some of them are not.

The design and control scheme of the drum granulator-dryer normally force layered growth or coating and block coalescence or agglomeration. Sometimes the formation of undesirable agglomerates indicates a shift of granulation regime from layering to coalescence, which is not a normal case of operation and must be avoided.

Granule growth by spraying the slurry onto the previously formed seed is shown in Figure 3.

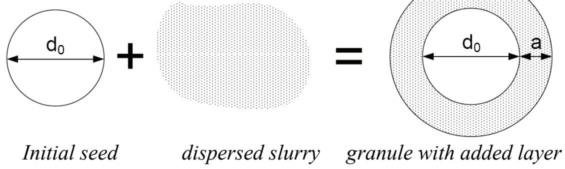


Figure 3: Granule growth by layering.

To model the layering phenomenon, the thickness of a new layer applied is determined by the diameter of the initial particle and the volume of the slurry applied. Assuming a spherical primary particle and a uniform distribution of all sprayed slurry applied onto the particle, the volume of the added layer V_l is calculated from the difference in the volumes of the layered particle and the initial one:

The thickness of the applied layer:

$$a = \frac{1}{2\pi}\sqrt[3]{\pi^2(\pi d_0^3 + 6V_l)} - \frac{1}{2}d_0 \qquad (1)$$

here d_0 – initial width of granule (seed), a – thickness of the applied layer.

The explicit mass and energy balance model with its wide and quite complex mathematical and physical features is beyond the scope of this paper. Hence, the following is the simplified version of the model developed.

The overall mass balance inside the granulator in liquid phase:

$$\frac{dM_L}{dt} = F_{L,in} - F_{L,out} - F_e - \dot{m}_c \qquad (2)$$

here M_L – accumulated mass of liquid solution, $F_{L,in}$ – flow of liquid solution into the granulator, $F_{L,out}$ – flow of liquid solution out of the granulator, F_e – flow of evaporated liquid solution, m_c – mass of crystallized solution (solid material).

The overall mass balance inside the granulator in solid phase:

$$\frac{dM_S}{dt} = F_{S,in} - F_{S,out} + \dot{m}_c + \dot{m}_g - \dot{m}_{att} \qquad (3)$$

here M_S – accumulated mass of solid material, $F_{S,in}$ – flow of solids into the granulator, $F_{S,out}$ – flow of solids out of the granulator, m_g – mass due to growth, $\mathrm{m}_{\mathrm{att}}$ – mass due to attrition.

The overall energy balance inside the granulator:

$$\frac{dE}{dt} = \dot{E}_{in} + \dot{E}_f + \dot{E}_r - \dot{E}_e - \dot{E}_l - \dot{E}_{out} \qquad (4)$$

here E – overall energy, E_{in} – energy provided into the granulator, E_{out} – energy removed from the granulator, E_f – energy due to gas furnace action, E_r – energy of reaction heat, E_e – energy for moisture evaporation, E_l – loss of energy from the granulator to environment.

The model presented is placed in stochastic background, which can better suit the growth kinetics, heat and mass transfer phenomena that actually happen in the real plant, with addition of uncertainty and plant equipment properties.

This section has presented only a part of the general model, which is in nature a grey box. Complementary models from measured process data have been also built and expert information used to enrich the model presented.

3.2 Statistical Analysis for Modeling

Nowadays it is possible to measure, store and retrieve process sensors' data and afterwards perform statistical analysis to "mine" some knowledge. For this purpose, descriptive and inferential statistics need to be used.

Taking different combinations of data sets of the essential process variables, the following results have been obtained:

(1) Scatter plots of the parameters.

(2) Reduced linear correlation matrix with entries of defined correlation degree (used for fast determination of parameter combinations which have a strong linear correlation).

(3) Linear models of the first order polynomial (application of stepwise regression, which is a technique for choosing variables, i.e. terms, to include in a multiple regression model).

(4) Residuals, confidence intervals of parameters, t-statistic, p-value, R^2 calculated for the generated linear models.

(5) Plots of cross-correlation function for probable lead/lag determination.

Figure 4 presents the fragment correlation and regression analysis of two process parameters (*3* and *7*).

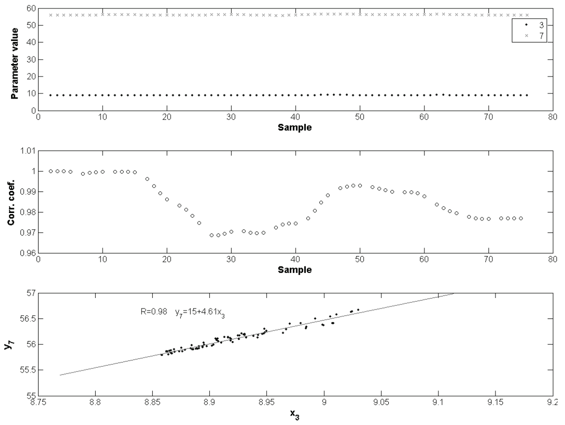


Figure 4: Results of the regression analysis of process parameters.

These findings are significant for determination of process parameters and their relationship. They can also provide additional knowledge for the plant diagnostics. A more detailed statistical analysis of granulation process can be found in (Valiulis and Simutis, 2007).

Mere statistical analysis is rarely helpful. Some heuristic knowledge should be also applied to make it work.

3.3 Knowledge-based Modeling

Complex multiscale process systems which are difficult to model properly (such as granulation) require a combination of various analytical and heuristic techniques. Effective solutions are often based on information from heterogeneous knowledge sources. One of them is knowledge-based systems built on the methods and techniques of Artificial Intelligence.

The expert knowledge of the process is an invaluable source of knowledge, especially, when there is a lack of reliable physical description and suitable measurement equipment. Rule-based expert systems use "if…then…" rules to represent human expert knowledge, which is often a mix of theoretical knowledge, heuristics derived from experience, and special-purpose rules for dealing with abnormal situations (Shang, 2004).

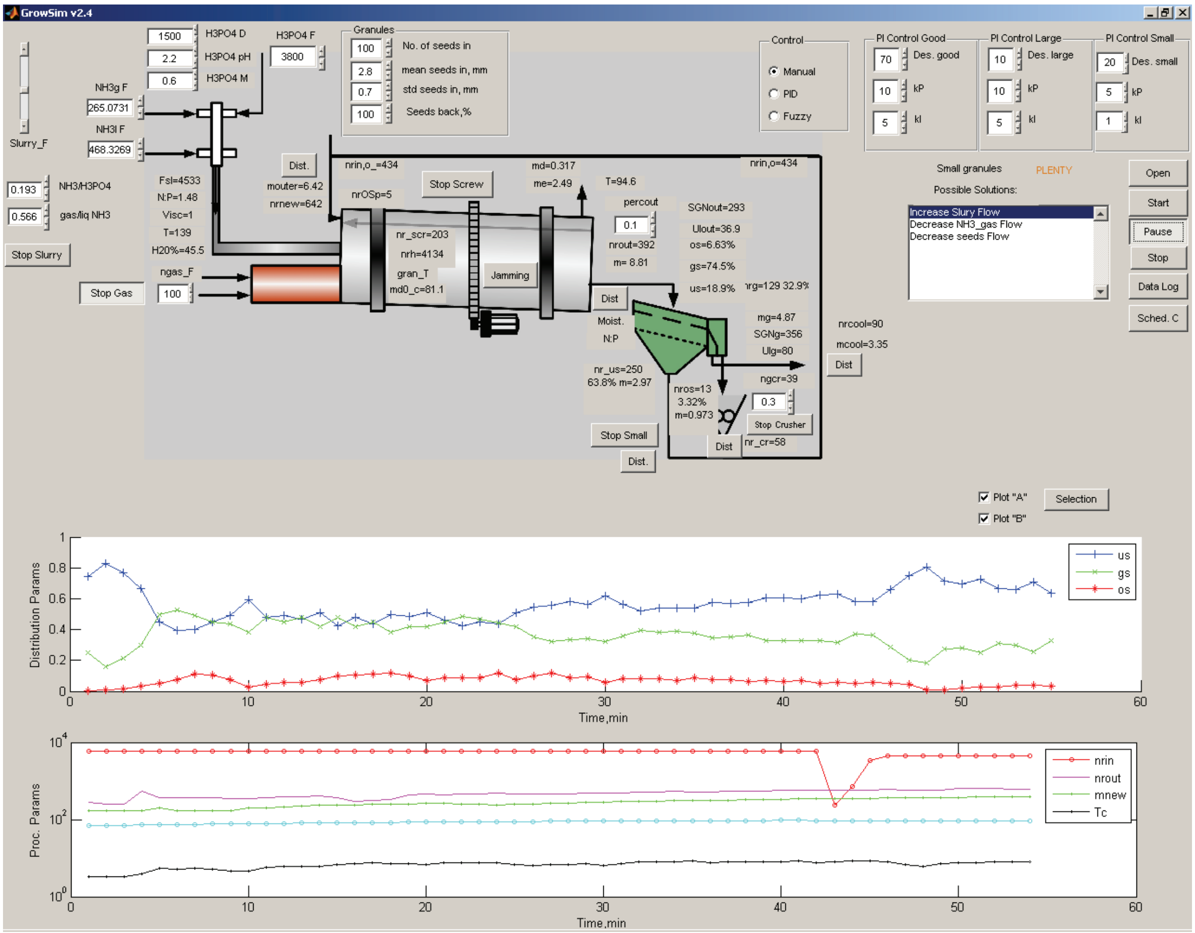
An example of the "if…then…" rule of new seed formation inside the drum granulator-dryer is presented as follows:

If *granule curtain in the spray zone is poor* and *gas temperature in the spray zone is high*, then *new small nuclei formation rate is high*.

In the proposed modeling approach, the expert knowledge is represented by the rule set. The rules involve variables such as "poor", "high", dealing with fuzziness, which is very common in real world problems. Unlike conventional expert systems, which are mainly symbolic reasoning engines, fuzzy expert systems are oriented toward numerical processing (Hemmer, 2008). These principles can be applied for the future development of the granulation process model and simulator for automated guidance and diagnostic purposes.

4 SIMULATOR

Increasing capabilities of computer hardware and software ensure the incorporation of complex knowledge (models) represented by differential and algebraic equations, measured process data, process experts' information, etc. But to be of use for the day-to-day work of the engineer these models have to become more user friendly, than the one that the scientist is dealing with (Ihlow et al., 2004). A new "GrowSim" simulation package for granulation process modeling and simulation is under


GrowSim v2.4

Slurry_F

1500 H3PO4 D
2.2 H3PO4 pH
0.6 H3PO4 M
H3PO4 F 3800

NH3g F 265.0731
NH3l F 468.3269

0.193 NH3/H3PO4
0.566 gas/liq NH3

Granules:
100 No. of seeds in
2.8 mean seeds in, mm
0.7 std seeds in, mm
100 Seeds back,%

Control: ● Manual ○ PID ○ Fuzzy

PI Control Good: 70 Des. good; 10 kP; 5 kI
PI Control Large: 10 Des. large; 10 kP; 5 kI
PI Control Small: 20 Des. small; 5 kP; 1 kI

Small granules PLENTY

Possible Solutions:
Increase Slury Flow
Decrease NH3_gas Flow
Decrease seeds Flow

Fsl=4533
N:P=1.48
Visc=1
T=139
H20%=45.5
ngas_F 100

mouter=6.42
nrnew=642
nrin,o_=434
nrOSp=5
nr_scr=203
nrh=4134
gran_T
md0_c=81.1
md=0.317
me=2.49
T=94.6
percout 0.1
nrout=392
m= 8.81
nrin,o=434
SGNout=293
Ulout=36.9
os=6.63%
gs=74.5%
us=18.9%
nrg=129 32.9%
mg=4.87
SGNg=356
Ulg=80
nrcool=90
mcool=3.35
nr_us=250
63.8% m=2.97
nros=13
3.32%
m=0.973
ngcr=39
0.3
nr_cr=58

☑ Plot "A"
☑ Plot "B"

Figure 5: Graphical user interface of “GrowSim” simulator.

development to realize this concept. The simulator is intended to be used by novice process operators to improve their skills in process control and to acquire knowledge of underlying mechanisms. The graphical user interface (Figure 5) has been built to mimic the process control environment available to the process operator in the real plant, with important additional information provided.

The simulation environment is composed of sections where the operator can change the manipulative process parameters, observe the current or past output parameters, get some advice on how the process in the current state should be controlled by the skilled operator. The manual or automatic process control modes are available. The operator can take a challenge to manage the process by hand or leave all or part of the job to PID or Fuzzy controller. The simulation time is much shorter compared to the real process and can be easily adjusted if the computing resources are sufficient. It is important to note that simulation can be paused at any moment, to make it possible to weigh one’s decisions. The main process parameters are stored and can be observed during the simulation or even later. The simulator is provided with routine to compare the simulated and real measured process data.

5 VALIDATION EXAMPLES

Some experiments have been carried out to validate model performance against measured plant data in prediction of the granule size distribution. Experimental data have been obtained using particle image analysis system.

Figure 6 presents the impact of slurry feed rate on the cumulative granule size distribution.

In *phase A* the process is kept in some steady state, with the median granule size nearly 2.65 mm. In *case B* the slurry feed rate is increased approximately by 35 %. In this situation the granule median size is nearly 2.9 mm. Change in the slurry flow rate alters the granule growth rate. Increase in the slurry flow rate raises the granule growth rate and a shift of cumulative granule size distribution to the right is observed.

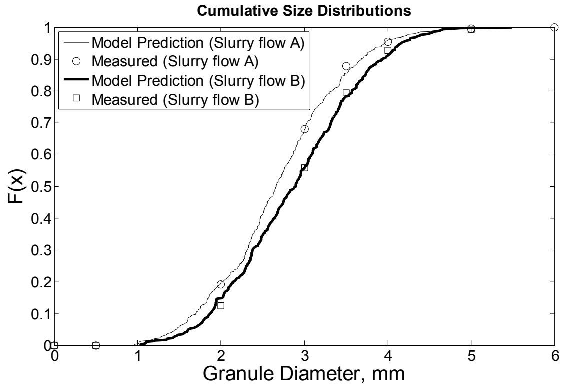


Figure 6: Impact of the change of the slurry flow rate on the cumulative granule size distribution.

Figure 7 presents the measured and simulated cumulative granule size distributions of initial seeds fed to the granulator (mean size is about 2.75 mm) and granules flowing out of the granulator (mean size is about 3.15 mm) in steady state.

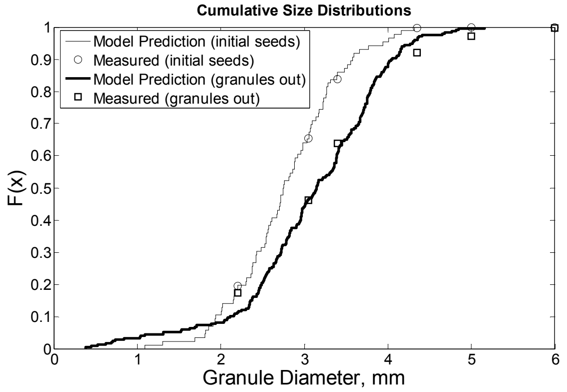


Figure 7: Cumulative size distributions of initial seed flow and granule flow out of the granulator.

6 CONCLUSIONS

An industrial DAP fertilizer granulation circuit, including drum, sieves, crusher, transportation system, has been modeled using basic physical principles such as growth kinetics, mass and heat transfer. Statistical analysis and system identification procedures have been performed for the estimation of unknown model parameters. As an extension to the model, additional information has been extracted from the process experts to define unmeasured parameters and assess some equipment properties. The whole model has been implemented and simulation executed using "GrowSim" simulator in the MATLAB environment.

Some model validation procedures have been performed and the results appear to be in fair accordance with the plant measured data. The findings, presented in Figures 6 and 7, show some kind of mismatch, but still can be treated satisfactory.

The current and future research is focused on further model development, implementation and testing of different plant control modes such as PID, Fuzzy and model predictive control. The primary results demonstrate the need of combination of the aforementioned control methods for a robust process control.

REFERENCES

Degreve J., Baeyens J., Van de Velden M., De Laet S. (2006). Spray-agglomeration of NPK-fertilizer in a rotating drum granulator. *Powder Technology,* vol. 163, p. 176-183.

Findlay W.P., Peck G.R., Morris K.R. (2005). Determination of Fluidized Bed Granulation End Point Using Near-Infrared Spectroscopy and Phenomenological Analysis. *Journal of Pharmaceutical Sciences*, vol. 94, p. 604-612.

Hemmer M.C (2008). *Expert Systems in Chemistry Research*. CRC Press.

Ihlow M., Drechsler J., Peglow M., Henneberg M., Mörl L. (2004). A New Comprehensive Model and Simulation Package for Fluidized Bed Spray Granulation Processes. *Chemical Engineering and Technology,* vol. 27, p. 1139-1143.

Salmon A.D., Hounslow M.J., Seville J.P.K. (2007). *Handbook of Powder Technology, vol. 11. Granulation*. Elsevier.

Shang Yi (2004). *Expert Systems. The Electrical Engineering Handbook (Section 5)*. Elsevier Academic Press, p. 367-377.

Valiulis G., Simutis R. (2007). Application of Regression Analysis for Modelling of Granule Size Distribution. *Proceedings of International Conference "Electrical and Control Technologies – 2007",* p. 62-67.

Wang F.Y, Cameron I.T., (2002). Review and future directions in the modelling and control of continuous drum granulation. *Powder Technology,* vol. 124, p. 238-253.

FILTERING AND COMPRESSION OF STOCHASTIC SIGNALS UNDER CONSTRAINT OF VARIABLE FINITE MEMORY

Anatoli Torokhti and Stan Miklavcic
University of South Australia, Adelaide, Australia
anatoli.torokhti@unisa.edu.au, stan.miklavcic@unisa.edu.au

Keywords: Optimal filtering, Wiener filter.

Abstract: We study a new technique for optimal data compression subject to conditions of causality and different types of memory. The technique is based on the assumption that certain covariance matrices formed from observed data, reference signal and compressed signal are known or can be estimated. In particular, such an information can be obtained from the known solution of the associated problem with no constraints related to causality and memory. This allows us to consider two separate problems related to compression and de-compression subject to those constraints. Their solutions are given and the analysis of the associated errors is provided.

1 INTRODUCTION

A study of data compression methods is motivated by the necessity to reduce expenditures incurred with the transmission, processing and storage of large data arrays. While the topics have been intensively studied (see e.g. (S. Friedland, 2006), (Jolliffe, 1986), (Hua and Nikpour, 1999), (Hua and Liu, 1998), (A. Torokhti, 2001), (Torokhti and Howlett, 2007), (T. Zhang, 2001)), a number of related fundamental questions are still open. One of them concerns specific restrictions associated with different types of causality and memory.

First Motivation: Causality and Memory. Data compression techniques mainly consist of three operations, compression itself, de-noising and de-compression (or reconstruction) of the compressed data. Each operation is implemented by a special filter. In reality, a value of the output of such a filter at time t_k is determined from a 'fragment' of its input defined at times $t_k, t_{k-1}, \ldots, t_{k-q}$. In other words, in practice both operations are subject to the conditions of causality and memory.

Our first motivation comes from a real-time signal processing. This implies that the filters we propose should be causal with variable finite memory.

Second Motivation: Reformulation of the Problem. Let (Ω, Σ, μ) be a probability space, where $\Omega = \{\omega\}$ is the set of outcomes, Σ a σ–field of measurable subsets in Ω and $\mu : \Sigma \rightarrow [0,1]$ an associated probability measure on Σ with $\mu(\Omega) = 1$.

In an informal way, the data compression problem we consider can be expressed as follows. Let $\mathbf{y} \in L^2(\Omega, \mathbb{R}^n)$ be observable data and $\mathbf{x} \in L^2(\Omega, \mathbb{R}^m)$ be a reference signal that is to be estimated from $\mathbf{y}$ in such a way that, (a) the data $\mathbf{y}$ should be compressed to a 'shorter' vector $\mathbf{z} \in L^2(\Omega, \mathbb{R}^r)$[1] with $r < \min\{m, n\}$ and (b) $\mathbf{z}$ should be de-compressed (reconstructed) to a signal $\tilde{\mathbf{x}} \in L^2(\Omega, \mathbb{R}^m)$ that is 'close' to $\mathbf{x}$ in some appropriate sense. Both operations should be *causal* and have *variable finite memory*. In this paper, the term 'close' is used with respect to the minimum of the norm (2) of the difference between $\mathbf{x}$ and $\tilde{\mathbf{x}}$.

The problem can be formulated in several alternate ways.

The first way is as follows. Let $\mathcal{B} : L^2(\Omega, \mathbb{R}^n) \rightarrow L^2(\Omega, \mathbb{R}^r)$ signify compression so that $\mathbf{z} = \mathcal{B}(\mathbf{y})$ and let $\mathcal{A} : L^2(\Omega, \mathbb{R}^r) \rightarrow L^2(\Omega, \mathbb{R}^m)$ designate data de-compression, *i.e.*, $\tilde{\mathbf{x}} = \mathcal{A}(\mathbf{z})$. We suppose that $\mathcal{B}$ and $\mathcal{A}$ are linear operators defined by the relationships

$$[\mathcal{B}(\mathbf{y})](\omega) = B[\mathbf{y}(\omega)] \quad \text{and} \quad [\mathcal{A}(\mathbf{z})](\omega) = A[\mathbf{z}(\omega)] \tag{1}$$

where $B \in \mathbb{R}^{n\times r}$ and $A \in \mathbb{R}^{r\times m}$. In the remainder of

[1] Components of $\mathbf{z}$ are often called *principal components* (Jolliffe, 1986).

this paper we shall use the same symbol to represent both the linear operator acting on a random vector and its associated matrix.

We define the norm to be

$$\|\mathbf{x}\|_\Omega^2 = \int_\Omega \|\mathbf{x}(\omega)\|_2^2 d\mu(\omega) \tag{2}$$

where $\|\mathbf{x}(\omega)\|_2$ is the Euclidean norm of $\mathbf{x}(\omega)$. Let us denote by $J(A,B)$, the norm of the difference between $\mathbf{x}$ and $\tilde{\mathbf{x}}$, constructed by A and B:

$$J(A,B) = \|\mathbf{x} - (A \circ B)(\mathbf{y})\|_\Omega^2. \tag{3}$$

The problem is to find $B^0 : L^2(\Omega,\mathbb{R}^n) \to L^2(\Omega,\mathbb{R}^r)$ and $A^0 : L^2(\Omega,\mathbb{R}^r) \to L^2(\Omega,\mathbb{R}^m)$ such that

$$J(A^0,B^0) = \min_{A,B} J(A,B) \tag{4}$$

subject to conditions of causality and variable finite memory for A and B. The problem consists of two unknowns, A and B.

A second way to formulate the problem, that avoids a difficulty associated with the two unknowns, is as follows. Let $\mathcal{F} : L^2(\Omega,\mathbb{R}^n) \to L^2(\Omega,\mathbb{R}^m)$ be a linear operator defined by

$$[\mathcal{F}(\mathbf{y})](\omega) = F[\mathbf{y}(\omega)] \tag{5}$$

where $F \in \mathbb{R}^{n\times m}$. Let rank $F = r$ and

$$J(F) = \|\mathbf{x} - \mathcal{F}(\mathbf{y})\|_\Omega^2.$$

Find $\mathcal{F}^0 : L^2(\Omega,\mathbb{R}^n) \to L^2(\Omega,\mathbb{R}^m)$ such that

$$J(F^0) = \min_F J(F) \tag{6}$$

subject to

$$\text{rank}\, F \le \min\{m,n\} \tag{7}$$

and conditions of causality and variable finite memory for F. Unlike (4), the problem (6)–(7) has only one unknown.

2 STATEMENT OF THE PROBLEM

The basic idea of our approach is as follows.

Let $\mathbf{x} \in L^2(\Omega,\mathbb{R}^m)$, $\mathbf{y} \in L^2(\Omega,\mathbb{R}^n)$ and $\mathbf{z} \in L^2(\Omega,\mathbb{R}^r)$, and let A and B be defined as (1) below. Here, $\mathbf{z}$ is a compressed version of $\mathbf{x}$. We assume that information about vector $\mathbf{z}$ in the form of associated covariance matrices can be obtained, in particular, from the known solution (Torokhti and Howlett, 2007) of problem (6)-(7) *with no constraints associated with causality and memory.*

In this paper, the data compression problem *subject to conditions of causality and memory* is stated in the form of two separate problems, (8) and (10) formulated below.

We use the following notation: $\mathcal{M}(r,n,\eta_B)$ is a set of causal $r\times n$ matrices B with a so-called *complete* variable finite memory η_B. The notation $\mathcal{M}(m,r,\eta_A)$ is similar.

Consider

$$J_1(B) = \|\mathbf{z} - B(\mathbf{y})\|_\Omega^2.$$

Let B^0 be such that

$$J_1(B^0) = \min_B J_1(B) \quad \text{subject to } B \in \mathcal{M}(r,n,\eta_B). \tag{8}$$

We write $\mathbf{z}^0 = B^0(\mathbf{y})$. Next, let

$$J_2(A) = \|\mathbf{x} - A(\mathbf{z}^0)\|_\Omega^2 \tag{9}$$

and let A^0 be such that

$$J_2(A^0) = \min_A J_2(A) \quad \text{subject to } A \in \mathcal{M}(m,r,\eta_A). \tag{10}$$

We denote $\mathbf{x}^0 = A^0(\mathbf{z}^0)$.

The problem considered in this paper is to find operators B^0 and A^0 that satisfy minimization criteria (8) and (10), respectively.

The major differences between the above statement of the problem and the statements considered below are as follows.

First, A and B should be causal with variable finite memory.

Second, it is assumed that certain covariance matrices formed from $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$ are known or can be estimated. In particular, such information can be obtained from the known solution (Torokhti and Howlett, 2007) of problem (6)-(7) with no constraints associated with causality and memory. We note that such an assumption does not look too restrictive in comparison with the assumptions used in the associated methods (Hua and Nikpour, 1999)–(Torokhti and Howlett, 2007).

Consequently and thirdly, we represent the initial problem in the form of a concatenation of two new separate problems (8) and (10).

3 MAIN RESULTS

Let $\tau_1 < \tau_2 < \cdots < \tau_n$ be time instants and $\alpha,\beta,\vartheta : \mathbb{R} \to L^2(\Omega,\mathbb{R})$ be continuous functions. Suppose $\alpha_k = \alpha(\tau_k)$, $\beta_k = \beta(\tau_k)$ and $\vartheta_k = \vartheta(\tau_k)$ are real-valued random variables having finite second moments. We write $\mathbf{x} = [\alpha_1, \alpha_2, \ldots, \alpha_m]^T$ $\mathbf{y} = [\beta_1, \beta_2, \ldots, \beta_n]^T$ and $\mathbf{z} = [\vartheta_1, \ldots, \vartheta_r]^T$.

Let $\tilde{\mathbf{z}}$ be a compressed form of data $\mathbf{y}$ defined by $\tilde{\mathbf{z}} = B(\mathbf{y})$ with $\tilde{\mathbf{z}} = [\tilde{\vartheta_1}, \ldots, \tilde{\vartheta_r}]^T$, and $\tilde{\mathbf{x}}$ be a decompression of $\tilde{\mathbf{z}}$ defined by $\tilde{\mathbf{x}} = A(\tilde{\mathbf{z}})$ with $\tilde{\mathbf{x}} = [\tilde{\alpha_1}, \ldots, \tilde{\alpha_m}]^T$.

In many applications[2], to obtain $\tilde{\vartheta_k}$ for $k = 1, \ldots, r$, it is necessary for B to use only a limited number of input components, $\eta_{B_k} = 1, \ldots, r$. A number of such input components η_{B_k} is here called a *kth local memory* for B.

To define a notation of memory for the compressor B, we use parameters p and g which are positive integers such that $1 \le p \le n$ and $n - r + 2 \le g \le n$.

Definition 1. *The vector $\eta_B = [\eta_{B_1}, \ldots, \eta_{B_r}]^T \in \mathbb{R}^r$ is called* a variable memory *of the compressor B. In particular, η_B is called a* complete *variable memory if $\eta_{B_1} = g$ and $\eta_{B_k} = n$ when $k = n - g + 1, \ldots, n$. Here, p relates to the last possible nonzero entry in the bottom row of B and g relates to the last possible nonzero entry in the first row.*

The notation $\eta_A = [\eta_{A_1}, \ldots, \eta_{A_m}]^T \in \mathbb{R}^m$ has a similar meaning for the de-compressor A, *i.e.*, η_A is *a variable memory* of the de-compressor A. Here, η_{A_j} is the *jth local memory* of A.

The parameters q and s, which are positive integers such that $1 \le q \le r$ and $2 \le s \le m$, are used below to define two types of memory for A.

Definition 2. *Vector η_A is called a* complete *variable memory of the de-compressor A if $\eta_{A_1} = q$ and $\eta_{A_j} = r$ when $j = s + r - 1, \ldots, m$. Here, q relates to the first possible nonzero entry in the last column of A and s relates to the first possible nonzero entry in the first column.*

The memory constraints described above imply that certain elements of the matrices $B = \{b_{ij}\}_{i,j=1}^{r,n}$ and $A = \{a_{ij}\}_{i,j=1}^{m,r}$ must be set equal to zero. In this regard, for matrix B with $r \le p \le n$, we require that

$$b_{i,j} = 0$$

$$\text{if} \quad j = p - r + i + 1, \ldots, n,$$

$$\text{for} \left\{ \begin{array}{l} p = r, \ldots, n-1, \\ i = 1, \ldots, r \end{array} \right. \quad \text{and} \quad \left\{ \begin{array}{l} p = n, \\ i = 1, \ldots, r-1, \end{array} \right.$$

and, for $1 \le p \le r - 1$, it is required that

$$b_{i,j} = 0$$

$$\text{if} \left\{ \begin{array}{l} i = 1, \ldots, r - p, \\ j = 1, \ldots, n, \end{array} \right. \quad \text{and} \quad \left\{ \begin{array}{l} i = r - p + 1, \ldots, r, \\ j = i - r + p + 1, \ldots, n. \end{array} \right.$$

For matrix A with $r \le p \le n$, we require

$$a_{i,j} = 0 \tag{11}$$

if$j = q + i, \ldots, r$for$q = 1, \ldots, r-1, i = 1, \ldots, r - q$,

and, for $2 \le s \le m$, it is required that

$$a_{i,j} = 0$$

if $j = s + i, \ldots, r$ for $s = 1, \ldots, m,\ i = 1, \ldots, s + r - 1$,

The above conditions imply the following definitions.

Definition 3. *A matrix B satisfying the constraint (11)–(11) is said to be a causal operator with the complete variable memory $\eta_B = [g, g+1, \ldots, n]^T$. Here, $\eta_{B_k} = n$ when $k = n - g + 1, \ldots, n$. The set of such matrices is denoted by $\mathcal{M}_C(r, n, \eta_B)$.*

Definition 4. *A matrix A satisfying the constraint (11)–(11) is said to be a causal operator with the complete variable memory $\eta_A = [r - q + 1, \ldots, r]^T$. Here, $\eta_{A_j} = r$ when $j = q, \ldots, m$. The set of such matrices is denoted by $\mathcal{M}_C(m, r, \eta_A)$.*

3.1 Solution of Problems (8) and (10)

To proceed any further we shall require some more notation. Let

$$\langle \alpha_i, \beta_j \rangle = \int_\Omega \alpha_i(\omega) \beta_j(\omega) d\mu(\omega) < \infty, \tag{12}$$

$$E_{xy} = \{\langle \alpha_i, \beta_j \rangle\}_{i,j=1}^{m,n} \in \mathbb{R}^{m \times n},$$

$$\mathbf{y}_1 = [\beta_1, \cdots, \beta_{g-1}]^T, \quad \mathbf{y}_2 = [\beta_g, \cdots, \beta_n]^T, \tag{13}$$

$$\mathbf{z}_1 = [\vartheta_1, \ldots, \vartheta_{g-1}]^T \quad \text{and} \quad \mathbf{z}_2 = [\vartheta_g, \ldots, \vartheta_n]^T. \tag{14}$$

The pseudo-inverse matrix (Golub and Loan, 1996) for any matrix M is denoted by $M^\dagger$. The symbol $\mathbb{O}$ designates the zero matrix.

Lemma 1. (Torokhti and Howlett, 2007) *If we define $\mathbf{w}_1 = \mathbf{y}_1$ and $\mathbf{w}_2 = \mathbf{y}_2 - P_y \mathbf{y}_1$ where*

$$P_y = E_{y_1 y_2} E_{y_1 y_1}^\dagger + D_y (I - E_{y_1 y_1} E_{y_1 y_1}^\dagger) \tag{15}$$

with D_y an arbitrary matrix, then $\mathbf{w}_1$ and $\mathbf{w}_2$ are mutually orthogonal random vectors.

[2] Examples include computer medical diagnostics (Gimeno, 1987) and problems of bio-informatics (H. Kim, 2005).

Let us first consider problem (8) when B has the complete variable memory $\eta_B = [g, g+1, \ldots, n]^T$ (see Definition 3).

Let us partition B into four blocks K_B, L_B, S_{B1} and S_{B2} so that $B = \begin{bmatrix} K_B & L_B \\ S_{B1} & S_{B2} \end{bmatrix}$, where

$K_B = \{k_{ij}\} \in \mathbb{R}^{n_b \times (g-1)}$ is a rectangular matrix,
$L_B = \{\ell_{ij}\} \in \mathbb{R}^{n_b \times n_b}$ is a lower triangular matrix,
$S_{B1} = \{s_{ij}^{(1)}\} \in \mathbb{R}^{(r-n_b)\times(g-1)}$,
$S_{B2} = \{s_{kl}^{(2)}\} \in \mathbb{R}^{(r-n_b)\times n_b}$
are rectangular matrices, and $n_b = n - g + 1$.

We have $B(\mathbf{y}) = \begin{bmatrix} T_B(\mathbf{w}_1) + L_B(\mathbf{w}_2) \\ S_B(\mathbf{w}_1) + S_{B2}(\mathbf{w}_2) \end{bmatrix}$, where $T_B = K_B + L_B P_y$ and $S_B = S_{B1} + S_{B2} P_y$. Then

$$J_1(B) = J^{(1)}(T_B, L_B) + J^{(2)}(S_B, S_{B2}), \qquad (16)$$

where $J^{(1)}(T_B, L_B) = \|\mathbf{z}_1 - [T_B(\mathbf{w}_1) + L_B(\mathbf{w}_2)]\|_\Omega^2$, $J^{(2)}(S_B, S_{B2}) = \|\mathbf{z}_2 - [S_B(\mathbf{w}_1) + S_{B2}(\mathbf{w}_2)]\|_\Omega^2$. By analogy with Lemma 37 in (Torokhti and Howlett, 2007),

$$\min_{B \in \mathcal{M}(r,n,\eta_B)} J_1(B) = \min_{T_B, L_B} J^{(1)}(T_B, L_B) + \min_{S_B, S_{B2}} J^{(2)}(S_B, S_{B2}).$$

Therefore, problem (8) is reduced to finding matrices T_B^0, L_B^0, S_B^0 and S_{B2}^0 such that

$$J^{(1)}(T_B^0, L_B^0) = \min_{T_B, L_B} J^{(1)}(T_B, L_B) \qquad (17)$$

and

$$J^{(2)}(S_B^0, S_{B2}^0) = \min_{S_B, S_{B2}} J^{(2)}(S_B, S_{B2}). \qquad (18)$$

Taking into account the orthogonality of vectors $\mathbf{w}_1$ and $\mathbf{w}_2$ and working in analogy with the argument on pp. 348–352 in (Torokhti and Howlett, 2007), it follows that matrices S_B^0 and S_{B2}^0 are given by

$$S_B^0 = E_{z_2} E_{w_1 w_1}^\dagger + H_B(I - E_{w_1 w_1} E_{w_1 w_1}^\dagger) \qquad (19)$$

and

$$S_{B2}^0 = E_{z_2} E_{w_2 w_2}^\dagger + H_{B2}(I - E_{w_2 w_2} E_{w_2 w_2}^\dagger), \qquad (20)$$

where H_B and H_{B2} are arbitrary matrices.

Next, to find T_B^0 and L_B^0 we use the following notation.

For $r = 1, 2, \ldots, \ell$, let ρ be the rank of the matrix $E_{w_2 w_2} \in \mathbb{R}^{n_2 \times n_2}$ with $n_b = n - g + 1$, and let

$$E_{w_2 w_2}{}^{1/2} = Q_{w,\rho} R_{w,\rho} \qquad (21)$$

be the QR-decomposition for $E_{w_2 w_2}{}^{1/2}$ where $Q_{w,\rho} \in \mathbb{R}^{n_2 \times \rho}$ and $Q_{w,\rho}{}^T Q_{w,\rho} = I$ and $R_{w,\rho} \in \mathbb{R}^{\rho \times n_2}$ is upper trapezoidal with rank ρ. We write $G_{w,\rho} = R_{w,\rho}{}^T$ and use the notation $G_{w,\rho} = [g_1, \ldots, g_\rho] \in \mathbb{R}^{n_2 \times \rho}$ where $g_j \in \mathbb{R}^{n_2}$ denotes the j-th column of $G_{w,\rho}$. We also write $G_{w,s} = [g_1, \ldots, g_s] \in \mathbb{R}^{n_2 \times s}$ for $s \le \rho$ to denote the matrix consisting of the first s columns of $G_{w,\rho}$. For simplicity, let us denote this $G_s := G_{w,s}$. Next, let $\mathbf{e}_1{}^T = [1, 0, 0, 0, \ldots]$, $\mathbf{e}_2{}^T = [0, 1, 0, 0, \ldots]$, $\mathbf{e}_3{}^T = [0, 0, 1, 0, \ldots]$, etc. denote the unit row vectors irrespective of the dimension of the space.

Finally, any square matrix M can be written as $M = M_\Delta + M_\nabla$ where M_Δ is lower triangular and M_∇ is strictly upper triangular. We write $\|\cdot\|_F$ for the Frobenius norm.

Theorem 1. *Let* $B \in \mathcal{M}_C(r, n, \eta_B)$, i.e., *the compressor* B *is causal and has the complete variable memory* $\eta_B = [g, g+1, \ldots, n]^T$. *Then the solution to problem* (8) *is provided by the matrix* B^0, *which has the form* $B^0 = \begin{bmatrix} K_B^0 & L_B^0 \\ S_{B1}^0 & S_{B2}^0 \end{bmatrix}$, *where the blocks* $K_B^0 \in \mathbb{R}^{n_b \times (g-1)}$, $S_{B1}^0 \in \mathbb{R}^{(r-n_b)\times(g-1)}$ *and* $S_{B2}^0 \in \mathbb{R}^{(r-n_b)\times n_b}$ *are rectangular, and the block* $L_B^0 \in \mathbb{R}^{n_b \times n_b}$ *is lower triangular. These blocks are given as follows. The block* K_B^0 *is given by*

$$K_B^0 = T_B^0 - L_B^0 P_y \qquad (22)$$

with

$$T_B^0 = E_{z_1 w_1} E_{w_1 w_1}^\dagger + N_{B1}(I - E_{w_1 w_1} E_{w_1 w_1}^\dagger) \qquad (23)$$

where N_{B1} *is an arbitrary matrix. The block* $L_B^0 = \begin{bmatrix} \lambda_1^0 \\ \vdots \\ \lambda_{n_b}^0 \end{bmatrix}$, *for each* $s = 1, 2, \ldots, n_2$, *is defined by its rows*

$$\lambda_s^0 = \mathbf{e}_s{}^T E_{z_1 w_2} E_{w_2 w_2}{}^\dagger G_s G_s{}^\dagger + f_s{}^T (I - G_s G_s{}^\dagger) \qquad (24)$$

with $f_s{}^T \in \mathbb{R}^{1 \times n_2}$ *arbitrary. The blocks* S_{B1}^0 *and* S_{B2}^0 *are given by*

$$S_{B1}^0 = S_B^0 - S_{B2}^0 P_y \qquad (25)$$

and (20), *respectively. In* (25), S_B^0 *is presented by* (19). *The error associated with the compressor* B^0 *is given by*

$$\|\mathbf{z} - B^0 \mathbf{y}\|_\Omega^2 = \sum_{s=1}^{\rho} \sum_{j=s+1}^{n_2} |e_s{}^T E_{z_1 w_2} E_{w_2 w_2}{}^\dagger g_j|^2$$
$$+ \sum_{j=1}^{2} \|E_{z_j z_j}{}^{1/2}\|_F^2 - \sum_{i=1}^{2} \sum_{j=1}^{2} \|E_{z_i w_i} E_{w_j w_j}{}^{\dagger 1/2}\|_F^2. \qquad (26)$$

Let us now consider problem (10) when the decompressor A has the complete variable memory $\eta_A = [r - q + 1, \ldots, r]^T$ (see Definition 4).

In analogy with our partitioning of matrix B, we partition matrix A in four blocks K_A, L_A, S_{A1} and S_{A2} so that $A = \begin{bmatrix} K_A & L_A \\ S_{A1} & S_{A2} \end{bmatrix}$, where

$K_A = \{k_{ij}\} \in \mathbb{R}^{q \times (r-q)}$ is a rectangular matrix,
$L_A = \{\ell_{ij}\} \in \mathbb{R}^{q \times q}$ is a lower triangular matrix, and
$S_{A1} = \{s_{ij}^{(1)}\} \in \mathbb{R}^{(m-q)\times(r-q)}$,
$S_{A2} = \{s_{kl}^{(2)}\} \in \mathbb{R}^{(m-q)\times q}$
are rectangular matrices.

Let us partition $\mathbf{z}^0$ so that $\mathbf{z}^0 = \begin{bmatrix} \mathbf{z}_1^0 \\ \mathbf{z}_2^0 \end{bmatrix}$ with $\mathbf{z}_1^0 \in L^2(\Omega, \mathbb{R}^{r-q})$ and $\mathbf{z}_2^0 \in L^2(\Omega, \mathbb{R}^q)$. We also write

$$\mathbf{x}_1 = [\alpha_1 \ldots, \alpha_{r-q}]^T \quad \text{and} \quad \mathbf{x}_2 = [\alpha_{r-q+1}, \ldots, \alpha_m]^T,$$

and denote by $\mathbf{v}_1 \in L^2(\Omega, \mathbb{R}^{r-q})$ and $\mathbf{v}_2 \in L^2(\Omega, \mathbb{R}^q)$, orthogonal vectors according to Lemma 1 as

$$\mathbf{v}_1 = \mathbf{z}_1^0 \quad \text{and} \quad \mathbf{v}_2 = \mathbf{z}_2^0 - P_z \mathbf{z}_1^0,$$

where $P_z = E_{z_1 z_2} E_{z_1 z_1}^\dagger + D_z(I - E_{z_1 z_1} E_{z_1 z_1}^\dagger)$ with D_z an arbitrary matrix.

We write $G_{v,s} = [g_1, \ldots, g_s] \in \mathbb{R}^{q \times s}$ where $G_{v,s}$ is constructed from a QR-decomposition of $E_{v_2 v_2}{}^{1/2}$, in a manner similar to the construction of matrix $G_{w,s}$.

Furthermore, we shall define $G_s := G_{v,s}$.

Theorem 2. *Let $A \in \mathcal{M}_C(m, r, \eta_A)$, i.e. the de-compressor A is causal and has the complete variable memory $\eta_A = [r - q + 1, \ldots, r]^T$. Then the solution to problem* (10) *is provided by the matrix A^0, which has the form $A^0 = \begin{bmatrix} K_A^0 & L_A^0 \\ S_{A1}^0 & S_{A2}^0 \end{bmatrix}$, where the blocks $K_A^0 \in \mathbb{R}^{q \times (r-q)}$, $S_{A1}^0 \in \mathbb{R}^{(m-q)\times(r-q)}$ and $S_{A2}^0 \in \mathbb{R}^{(m-q)\times q}$ are rectangular, and the block $L_A^0 \in \mathbb{R}^{q \times q}$ is lower triangular. These blocks are given as follows. The block K_A^0 is given by*

$$K_A^0 = T_A^0 - L_A^0 P \tag{27}$$

with

$$T_A^0 = E_{x_1 v_1} E_{v_1 v_1}^\dagger + N_{A1}(I - E_{v_1 v_1} E_{v_1 v_1}^\dagger) \tag{28}$$

where N_1 is an arbitrary matrix. The block $L_A^0 = \begin{bmatrix} \lambda_1^0 \\ \vdots \\ \lambda_q^0 \end{bmatrix}$, for each $s = 1, 2, \ldots, q$, is defined by its rows

$$\lambda_s^0 = \mathbf{e}_s{}^T E_{x_1 v_2} E_{v_2 v_2}{}^\dagger G_s G_s{}^\dagger + f_s{}^T (I - G_s G_s{}^\dagger) \tag{29}$$

with $f_s{}^T \in \mathbb{R}^{1 \times q}$ arbitrary. The blocks S_{A1}^0 and S_{A2}^0 are given by

$$S_{A1}^0 = S_A^0 - S_{A2}^0 P, \;\; S_{A2}^0 = E_{x_2} E_{v_2 v_2}^\dagger + H_{A2}(I - E_{v_2 v_2} E_{v_2 v_2}^\dagger), \tag{30}$$

where

$$S_A^0 = E_{x_2} E_{v_1 v_1}^\dagger + H_A(I - E_{v_1 v_1} E_{v_1 v_1}^\dagger) \tag{31}$$

and H_{A2} and H_A are arbitrary matrices.

The error associated with the de-compressor A^0 is given by

$$\|\mathbf{x} - A^0 \mathbf{z}^0\|_\Omega^2 = \sum_{s=1}^{\rho} \sum_{j=s+1}^{q} |e_s{}^T E_{x_1 v_2} E_{v_2 v_2}{}^\dagger g_j|^2 \tag{32}$$

$$+ \sum_{j=1}^{2} \|E_{x_j x_j}{}^{1/2}\|_F^2 - \sum_{i=1}^{2} \sum_{j=1}^{2} \|E_{x_i v_i} E_{v_j v_j}{}^{\dagger 1/2}\|_F^2. \tag{33}$$

4 SIMULATIONS

The following simulations and numerical results illustrate the performance of the proposed approach.

Our filter $F^0 = A^0 B^0$ has been applied to compression, filtering and subsequent restoration of the reference signals given by the matrix $X \in \mathbb{R}^{256 \times 256}$. The matrix X represents the data obtained from an aerial digital photograph of a plant[3] presented in Fig. 1.

We divide X into 128 sub-matrices $X_{ij} \in \mathbb{R}^{m \times q}$ with $i = 1, \ldots, 16$, $j = 1, \ldots, 8$, $m = 16$ and $q = 32$ so that $X = \{X_{ij}\}$. By assumption, the sub-matrix X_{ij} is interpreted as q realizations of a random vector $\mathbf{x} \in L^2(\Omega, \mathbb{R}^m)$ with each column representing a realization. For each $i = 1, \ldots, 16$ and $j = 1, \ldots, 8$, observed data Y_{ij} were modelled from X_{ij} in the form

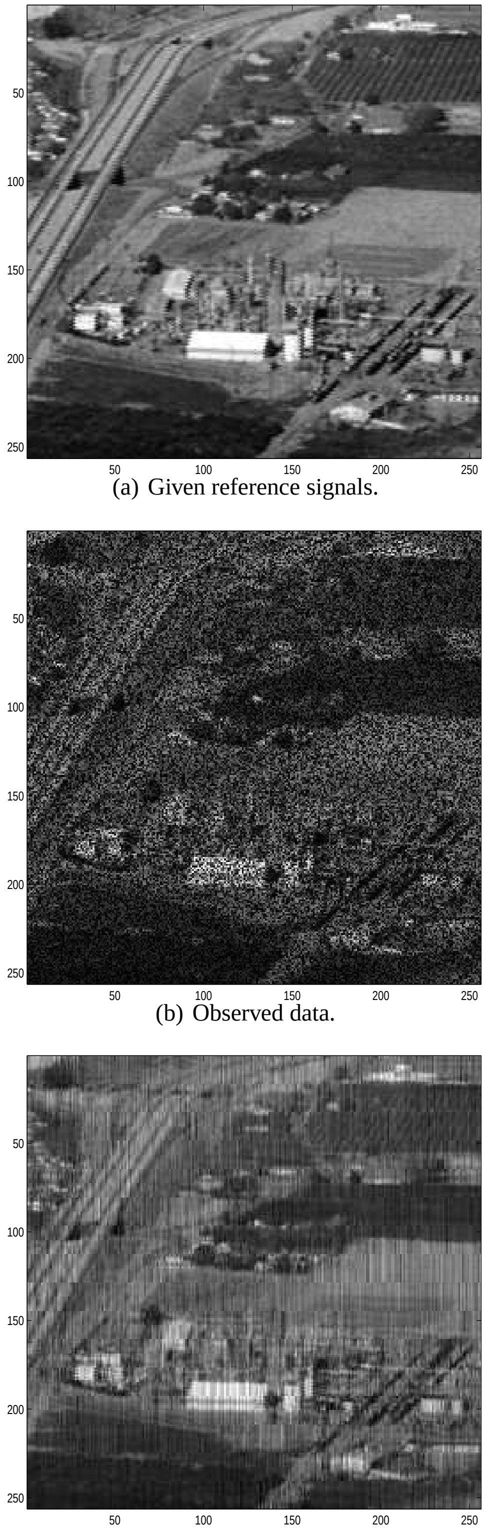
$$Y_{ij} = X_{ij} \bullet \mathtt{rand}(16, 32)_{(ij)}.$$

Here, $\bullet$ means the Hadamard product and $\mathtt{rand}(16, 32)_{(ij)}$ is a 16×32 matrix whose randomly-chosen elements are uniformly distributed in the interval $(0, 1)$.

The proposed filter F^0 has been applied to each pair $\{X_{ij}, Y_{ij}\}$. Each pair $\{X_{ij}, Y_{ij}\}$ was processed by compressors and de-compressors with the complete variable memory. We denote $B_C^0 = B^0$ and $A_C^0 = A^0$ for such a compressor and de-compressor determined by Theorems 1 and 2, respectively, so that

$$B_C^0 \in \mathcal{M}_T(r, n, \eta_B) \quad \text{and} \quad A_C^0 \in \mathcal{M}_C(m, r, \eta_A)$$

[3] The database is available in http://sipi.usc.edu/services/database/Database.html.



(a) Given reference signals.

(b) Observed data.

(c) Estimates of the reference signals by the filter F_C^0 with the complete variable memory.

Figure 1: Illustration of simulation results.

where $n = m = 16, r = 8, \eta_B = \{\eta_{Bk}\}_{k=1}^{16}$ with $\eta_{Bk} = \begin{cases} 12+k-1, & \text{if} \quad k = 1,\dots,4, \\ 16, & \text{if} \quad k = 5,\dots,16 \end{cases}$, and $\eta_A = \{\eta_{Aj}\}_{j=1}^{16}$ with $\eta_{Aj} = \begin{cases} 6+j-1, & \text{if} \quad j = 1,2, \\ 8, & \text{if} \quad k = 3,\dots,16 \end{cases}$. In this case, the optimal filter F^0 is denoted by F_C^0 so that $F_C^0 = A_C^0 B_C^0$. We write

$$J_C^0 = \max_{ij} \|X_{ij} - F_C^0 Y_{ij}\|^2$$

for a maximal error associated with the filter F_C^0 over all $i = 1,\dots,16$ and $j = 1,\dots,8$. The compression ratio was $c = 1/2$. We obtained $J_C^0 = 3.3123e+005$.

The results of simulations a are presented in Fig. 1 (a) - (c).

REFERENCES

A. Torokhti, P. H. (2001). Optimal fixed rank transform of the second degree. *IEEE Trans. Circuits & Syst., II, Analog & Digit. Signal Processing*, 48(3):309–315.

Gimeno, V. (1987). Obtaining the eeg envelope in real time: a practical method based on homomorphic filtering. *Neuropsychobiology*, 18:110–112.

Golub, G. and Loan, C. V. (1996). *Matrix Computation*. Johns Hopkins Univ. Press.

H. Kim, G.H. Golub, H. P. (2005). Missing value estimation for dna microarray gene expression data: local least squares imputation. *Bioinformatics*, 21:211–218.

Hua, Y. and Liu, W. Q. (1998). Generalized karhunen-loève transform. *IEEE Signal Process Letters*, 5:141–143.

Hua, Y. and Nikpour, M. (1999). Computing the reduced rank wiener filter by iqmd. *IEEE Signal Processing Letters*, 6(9):240–242.

Jolliffe, I. (1986). *Principal Component Analysis*. Springer Verlag.

S. Friedland, A. Niknejad, M. K. H. Z. (2006). Fast monte-carlo low rank approximations for matrices. *Proc. IEEE Conference SoSE*, pages 218–223.

T. Zhang, G. G. (2001). Rank-one approximation to high order tensors. *SIAM J. Matrix Anal. Appl.*, 23.

Torokhti, A. and Howlett, P. (2007). *Computational Methods for Modelling of Nonlinear Systems*. Elsevier.

PERIODIC DISTURBANCES REDUCTION IN THE CONTINUOUS CASTING PROCESS BY MEANS OF A MODIFIED SMITH PREDICTOR

Karim Jabri, Alain Mouchette, Bertrand Bèle
MC Department, ArcelorMittal Research, Maizières-Lès-Metz, France
{karim.jabri, alain.mouchette, bertrand.bele}@arcelormittal.com

Emmanuel Godoy, Didier Dumur
Control Department, Supélec, Gif-sur-Yvette, France
{emmanuel.godoy, didier.dumur}@supelec.fr

Keywords: Smith predictor, Continuous casting, Periodic disturbances, H-infinity framework.

Abstract: In the continuous casting process, various control strategies are used to reduce the mold level fluctuations which cause surface defects in the final product. This paper proposes a control structure able to improve the reduction of the bulging effect on the mold level. It is based on the Aström's modified Smith predictor scheme which presents the advantage that the setpoint response is decoupled from the disturbance rejection transfer function. H_∞ control theory is utilized to develop the controller of this second loop. Both the disturbances rejection and the robust stabilization are considered in this design. Effective tuning rules are also given. Simulation results confirm that the proposed design is more effective than the one based on the PID control law currently implemented in several real plants.

1 INTRODUCTION

In the steel industry, the continuous casting is the most used process to solidify the steel. Mold level control strategies are a key factor in ensuring the quality of the final product. Real implementation remains however complex because the controllers have to take into consideration the process uncertainties, the operating point changes and the disturbances affecting the casting. In order to lower the level fluctuations, several control theories have been applied in recent years. Some of them are already implemented at real plants. For example, an adaptive control law has been used to improve the mold level control accuracy (Kurokawa *et al.*, 1992). Matoba *et al*. applied the LQ control in the case of low speed casters (Matoba *et al.*, 1990). In the present paper, a new control design is proposed aiming at reducing the bulging effect on the mold within a guaranteed delay margin. The performances are compared to those of the currently implemented PID control law.

The paper is structured as follows. Next section describes the continuous casting machine, the phenomena disturbing the casting operations and presents the plant model and the PID control law implemented in the plants. Section III examines the Smith predictor and its modified version. Based on this one, the control structure designed using H_∞ framework is presented. Section IV validates in simulation the proposed structure showing its efficiency compared to PID.

2 CONTINUOUS CASTING MACHINE

2.1 Process Description

As shown in Figure 1, in the continuous casting machine, molten steel flows from the ladle through the tundish into the mold. The steel is solidified in the mold cooled by the water. A solidified shell is thus formed and continuously withdrawn out of the mold until the outlet of the machine where the steel fully solidified is cut into pieces used by different manufacturing processes.

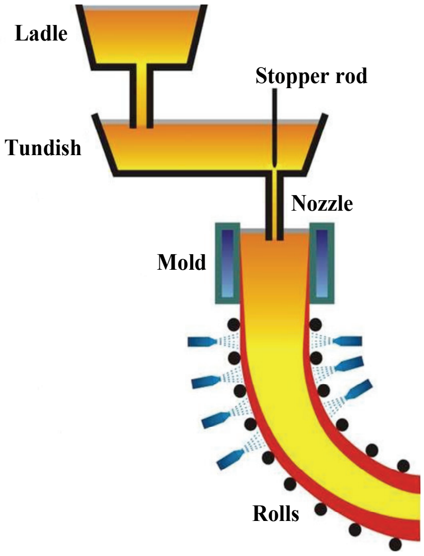


Figure 1: Continuous casting machine.

The steel level in the mold is a balance between the flows in and out of the mold. In order to regulate it, the actuator moves the stopper rod vertically to control the flow into the mold while the casting speed is kept constant. The controller uses also a sensor which measures only local level variations.

During casting operations, several disturbances occur and affect all the parts of the machine including the mold level regulation loop. The following two kinds of disturbances are dominant (Jabri *et al.*, 2008a).

2.2 Disturbances

The main disturbance considered here is the bulging which occurs between rolls due to increasing pressure inside the strand. Its profile is strongly affected by the roll pitch and lightly by the cooling conditions. Unsteady bulging generates important level fluctuations in the mold (Yoon *et al.*, 2002). Frequencies of this phenomenon appear to be in the range of 0.03-0.1Hz.

Other disturbances take place as the slow phase of clogging followed by a sudden unclogging that raises considerably the mold level (Thomas and Bai, 2001). There are also stationary surface waves of molten steel in the mold. Their frequencies depend on the mold width and are between 0.65 and 0.85Hz.

2.3 Plant model

Considering the description above and neglecting the level sensor dynamics, the plant model classically used for the design of the main control law is shown in Figure 2, with P^* the control input, P the stopper position, N the mold level, Q_{in} and Q_{out} the flow-rate into and out of the mold.

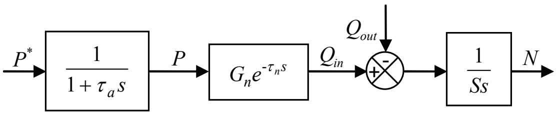


Figure 2: Plant model.

The parameters of the transfer functions appearing in the plant model are: τ_a the actuator time constant, G_n the stopper gain, τ_n the nozzle delay, S the mold section and s the Laplace variable. The process transfer function is thus given by:

$$H = \frac{G_n e^{-\tau_n s}}{Ss(1+\tau_a s)} = H_0 e^{-\tau_n s} \qquad (1)$$

In the plants, the mold level is often regulated by means of a PID controller which is not sufficient for bulging rejection. This current control strategy will be further used for comparison purposes. The tuning parameters are as follows, where the time constant of the derivative action filter is given by T_d / β :

$K = 0.38 \quad T_i = 9\,\mathrm{s} \quad T_d = 0.2\,\mathrm{s} \quad \beta = 10$

3 SMITH PREDICTOR CONTROL

3.1 Conventional Smith Predictor

The Smith predictor is widely used for the control of systems with time delays. It is a highly effective dead-time compensator especially for stable processes whose time delay is known (Figure 3).

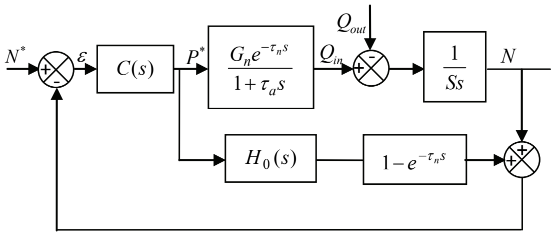


Figure 3: Conventional Smith predictor.

As shown in the equation below where H_0 is the delay free part of the plant model, the main advantage of the Smith predictor is that the delay is eliminated from the closed loop equation:

$$\frac{N}{N^*} = \frac{C(s)H_0(s)}{1 + C(s)H_0(s)} e^{-\tau_n s} \qquad (2)$$

If the flow out of the mold Q_{out} is equal to zero, the steady state error for a constant setpoint is equal to zero too because the open loop contains an integral term. However, the steady state error imposed by a constant flow out of the mold considered as a disturbance is not equal to zero because at low frequencies, its Laplace transform is given by:

$$\varepsilon_Q(s) = -N = \frac{-C(s)H_0(s)(e^{-\tau_n s} - 1) + 1}{Ss(1 + C(s)H_0(s))} Q_{out}$$
$$\underset{s \to 0}{\propto} \frac{\tau_n}{S} Q_{out} \qquad (3)$$

In order to avoid this problem, several authors have suggested modifications to the original Smith predictor. Lim *et al.,* 1990 proposes an extension based on the introduction of an additional feedback containing $G_n \tau_n$ in parallel with $H_0(s) - H(s)$. Although this structure cancels the steady state error, it does not allow users to tune the disturbances rejection which is a key factor in mold level control. The following paragraph describes the solution proposed by Aström to overcome this problem with the capability of shaping the frequency characteristics of the disturbances rejection (Chen *et al.*, 2007).

3.2 Aström's Modified Smith Predictor

In (Aström *et al.*, 1994), a two-degree of freedom modified Smith predictor is presented for first order integrative processes with dead time as shown in Fig. 4. The Astrom's Smith predictor decouples the disturbance response from the setpoint one and therefore can be independently optimized. Therefore, we can tune the performance of either setpoint tracking (through the transfer function $C(s)$) or disturbance rejection (through the transfer function $M(s)$) without affecting the other.

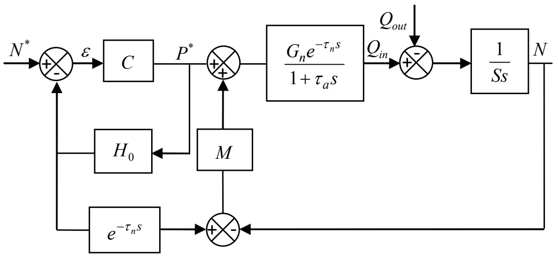


Figure 4: Aström's modified Smith predictor.

In this configuration, the setpoint response is given by:

$$\frac{N}{N^*} = \frac{C(s)H_0(s)}{1 + C(s)H_0(s)} e^{-\tau_n s} \qquad (4)$$

and the disturbance response is given by:

$$\frac{N}{Q_{out}} = \frac{-1}{Ss(1 + MH_0 e^{-\tau_n s})} \qquad (5)$$

In this work, the Aström's Smith predictor structure is used to reduce the influence of the bulging on the mold level.

In (Guanghui *et al.*, 2007), the proposed block diagram $M(s)$ is the following where $M_0(s)$ is the transfer function containing the tuning parameters:

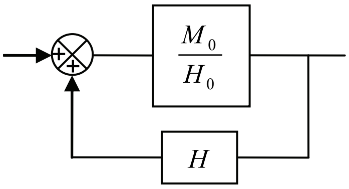


Figure 5: Proposed *M*(*s*) scheme.

Some tuning rules are given for $M_0(s)$ in order to eliminate the steady state error with a step disturbance. Unfortunately, this design does not improve the bulging rejection. Moreover, it uses the identification results of the gain and the delay and depends thus upon uncertainties on these two parameters.

Other simple forms of $M(s)$, e.g. first order function, have been investigated without success. In this paper, H_∞ control theory is used to shape the disturbance response by adjusting $M(s)$. Finally, the main controller $C(s)$ is chosen constant which is sufficient to tune the closed loop response time.

3.3 *M* Design using H_∞ Control Theory

For simplicity reasons, Figure 6 shows only the disturbance rejection loop. In order to achieve the foregoing specifications, a H_∞ control problem, described in Figure 7 and Figure 8, is established (Zhang *et al.*, 1991). A second disturbance W (which represents the standing waves actually) was added to the initial bulging rejection loop to be able to solve the H_∞ problem which requires several assumptions. In the proposed scheme, two weighting functions have been introduced. The first one W_1 is chosen to reduce the bulging effect on the level. The second one W_2 is tuned in order to achieve robust stability under delay changes and uncertainties.

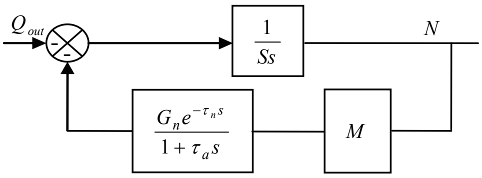


Figure 6: Disturbance rejection loop.

According to Figure 8, it comes:

$$\begin{cases} e_1 = W_1(s)B_{11}(s)Q_{out} + W_1(s)B_{12}(s)W \\ e_2 = W_2(s)B_{21}(s)Q_{out} + W_2(s)B_{22}(s)W \end{cases} \tag{6}$$

$$\text{with} \begin{cases} B_{11} = \frac{-1}{Ss}\frac{1}{(1+MH)} \quad B_{12} = \frac{1}{1-MH} \\ B_{21} = \frac{-G_n}{Ss(1+\tau_a s)}\frac{M}{(1+MH)} = \frac{-MH_0}{(1+MH)} \\ B_{22} = \frac{G_n}{(1+\tau_a s)}\frac{M}{(1-MH)} \end{cases}$$

Considering the state space formalism of the process described in Figure 7, the H_∞ control problem is formulated as follows:

$$\left\| \begin{pmatrix} W_1(s)B_{11}(s) & W_1(s)B_{12}(s) \\ W_2(s)B_{21}(s) & W_2(s)B_{22}(s) \end{pmatrix} \right\|_\infty \prec \gamma \tag{7}$$

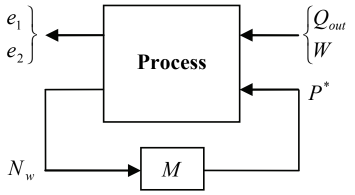


Figure 7: Standard H_∞ problem.

In order to approximate the time delay effect, the first order Pade function is used.

Since the bulging is described by a sinusoidal function with a frequency band between 0.03 and 0.1Hz, $B_{11}(s)$ should have a weak magnitude over this frequency range. First, W_1 is thus selected so that its gain is high over bulging frequencies and high enough on the low frequency band in order to eliminate the steady state error. In this work W_1^{-1} is chosen as a phase lead compensator:

$$W_1^{-1} = K_{w1}\frac{1+T_{w1}s}{1+a_{w1}T_{w1}s} \quad \text{with: } a_{w1} \prec 1 \tag{8}$$

Secondly, W_2 is tuned using the small gain theorem in order to achieve robust stability under delay changes. In fact, if the time delay changes less than $\Delta\tau_n$ (this upper bound is assumed to be known), the bulging rejection loop is stable if:

$$\left\| \frac{-HM}{1+MH}\Delta \right\|_\infty \prec 1 \tag{9}$$

with Δ a multiplicative uncertainty given by:

$$H_a = H(1+\Delta) \text{ and } \Delta = e^{\Delta\tau_n \cdot s} - 1 \tag{10}$$

Knowing that: $|H| = \left|H_0 e^{-\tau_n s}\right| = |H_0|$ (11)

(9) is equivalent to:

$$\left\| \frac{-H_0 M}{1+MH}\Delta \right\|_\infty = \|B_{21}\Delta\|_\infty \prec 1 \tag{12}$$

As Δ satisfies the following inequality:

$$|\Delta(j\omega)| \prec \left| \frac{2\Delta\tau_n \cdot j\omega}{1+\Delta\tau_n \cdot j\omega} \right| \tag{13}$$

W_2 is then chosen as:

$$|W_2(j\omega)| \succ \left| \frac{2\Delta\tau_n \cdot j\omega}{1+\Delta\tau_n \cdot j\omega} \right| \tag{14}$$

The two filters W_1 and W_2 should be calculated from equations (8) and (14). Finally, the H_∞ problem is solved using the Glover-Doyle's algorithm (Glover *et al.*, 1988).

4 SIMULATION RESULTS

The control structure designed in this way is tested by means of a mold level simulator developed with parameters issued from a real plant (Table 1). The previous tuning considers only the bulging rejection. The standing waves rejection was not explicitly taken into account.

Table 1: Plant model parameters.

Parameter	**Value**
τ_a	0.05s
τ_n	0.5s
G_n	10^6 mm^3/s/mm
S	1600×228mm^2
v	1.5m/min

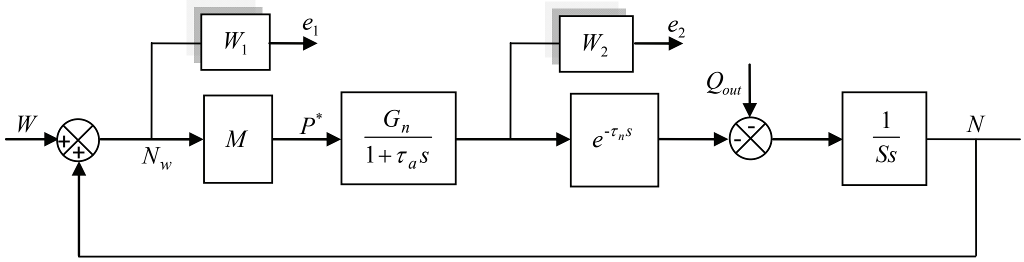


Figure 8: Block diagram of the proposed design.

The weight functions of the proposed design are:

$$W_1 = \frac{1+0.5s}{0.32+1.58s} \qquad W_2 = \frac{2.7s}{1+s}$$

W_2 was selected according to equation (14) in order to achieve a delay margin greater than the identified delay value (0.5s). In this case, the H_∞ controller is given by:

$$M_1 = \frac{27(s+20)(s+4)(s+1)(s+0.57)}{(s+482)(s+9.9)(s+1.2)(s+0.2)}$$

The stability and the robustness of the system controlled by the PID and the Aström's modified Smith predictor can be analyzed using the following diagrams. They show the control laws actions when the bulging occurs.

Figure 9 shows that the bulging rejection transfer function was improved with the modified Smith predictor. However, the steady state error is not equal to zero. In order to overcome this problem, the least of all the poles in M_1 was replaced by zero. Therefore, the new controller is given by (see Figure 10 for the new Bode diagram):

$$M_2 = \frac{27(s+20)(s+4)(s+1)(s+0.57)}{s(s+482)(s+9.9)(s+1.2)}$$

Figure 10 shows that the performances over the bulging frequency band are not modified. Those over lower frequencies are improved.

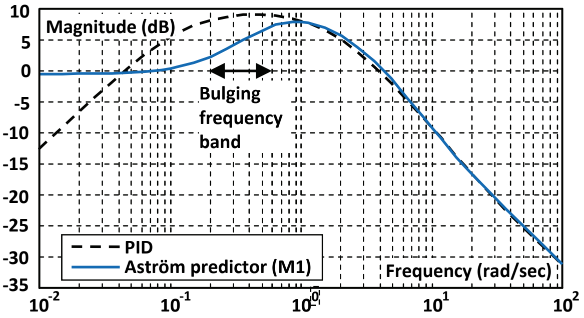


Figure 9: Bode diagram of the bulging rejection (case M_1).

Considering M_2, the main controller C was adjusted to set the closed loop response time ($C=1$). Figure 11 presents results obtained for a level variation of 10 mm.

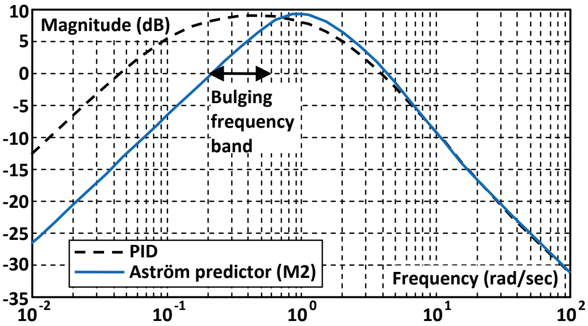


Figure 10: Bode diagram of the bulging rejection (case M_2).

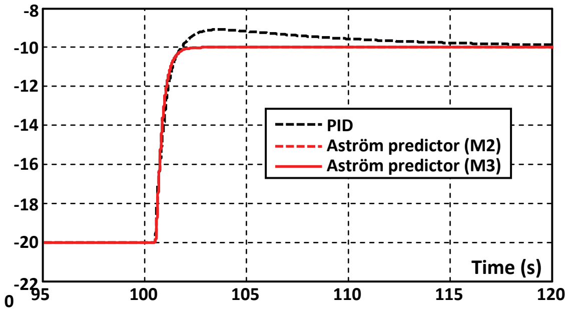


Figure 11: Mold level (mm).

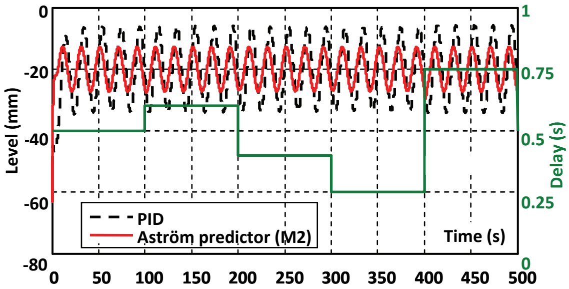


Figure 12: Mold level during bulging when delay changes.

Figure 12 shows the mold level when the delay changes during bulging whose frequency is 0.05Hz. Using M_2, the performances remain better than those of the PID.

M_2 can also be approached by a PID control law (see Figure 13 for the Bode diagrams) as follows:

$$M_3 = 0.51\left(1 + \frac{0.37}{s} + \frac{0.12s}{1 + 0.0025s}\right)$$

Finally, the performances of all the versions of the Aström's modified Smith predictor are summarized and compared with those of the PID in Table 2.

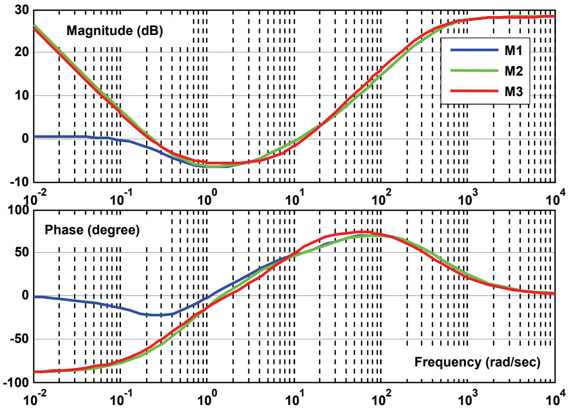


Figure 13: Bode diagrams of all the versions of Aström predictor.

Table 2: Performances of the proposed control laws.

Specifications	PID	Aström predictor		
		M_1	M_2	M_3
Cutoff frequency (rad/s)	1.06	1.3	1.3	1.41
Gain margin (dB)	8.7	10.1	9.9	10.4
Phase margin (°)	66	54	46	42
Delay margin (s)	1.1	0.75	0.61	0.52
$\max_{\omega\in[0.03\ 0.1\text{Hz}]} B_{11}(j\omega)$ (dB)	9	7.2	8.1	7.1
$\min_{\omega\in[0.03\ 0.1\text{Hz}]} B_{11}(j\omega)$ (dB)	8	2	-1.1	-0.6
Steady state error % outflow	0	small	0	0

5 CONCLUSIONS

This paper presents an effective method based on H_∞ control theory combined with the Aström's modified Smith predictor which enhances the disturbance rejection performance compared to the conventional Smith predictor. This one cannot indeed be utilized in the mold level control process since it leads to a steady state error as a response to a step disturbance.

Using simple tuning rules, the level error was reduced compared to the PID control with regards to robust stability. Moreover, this technique allows shaping the disturbance rejection independently from the closed loop response time which is not the case for PID. Further improvements may include additional features as the introduction of observers and feed-forward actions.

REFERENCES

Kurokawa, T., Kato, Y., Kondo T., 1992. Development of CC mold level adaptive control system. Current Advances in Materials and Processes, 5(2):354.

Matoba, Y., Yamamoto, T., Tozuda, M., Watanabe, T., Tomono, H., 1990. Instrumentation and control technology for supporting high-speed casting. In *9th PTD Process Technol. Conf. Proc.*

Jabri, K., Mouchette, A., Bèle, B., Godoy, E., Dumur, D., 2008a. Disturbances estimation for mold level control in the continuous casting process. In *5th International Conference on Informatics in Control, Automation and Robotics ICINCO*, Funchal, Portugal.

Yoon, U-S., Bang, I.-W., Rhee, J.H., Kim, S.-Y., Lee, J.-D., Oh, K.H., 2002. Analysis of mold level hunching by unsteady bulging during thin slab casting. ISIJ International, 42(10):1103-1111.

Thomas, B.G., Bai, H., 2001. Tundish nozzle clogging – application of computational models. In *18rd PTD Conf. Proc.*, Baltimore (US).

Lim, D.J., Jeong, H.S., Hong, D.H., 1990. A Study on the Mold Level Control of Continuous Casting System.

Chen, Y.D., Tung, P.C., Fuh, C.C., 2007. Modified Smith predictor scheme for periodic disturbance reduction in linear delay systems. *Journal of Process Control*, 17:799-804.

Aström, K.J., Hang, C.C., Lim, B.C., 1994. A new smith predictor for controlling a process with an integrator and long dead-time. *IEEE Trans. on Automatic Control*, 39(2):343-345.

Guanghui, Z., Feng, Q., Huihe, S., 2007. Robust tuning method for modified Smith predictor. *Journal of Systems Engineering and Electronics*, 18(1):89-94.

Zhang, F., Hosoe, S., Kouno, M., 1991. Synthesis of robust output regulators via H∞ control. In *13th SICE Symp. Dynamical Syst. Theory*.

Glover, K. et al, 1988. State space formula for all stabilizing controllers that satisfy an H∞ norm bound and relations to risk sensitivity. *Systems and Control Letters*, 11:167-172.

Jabri, K., Mouchette, A., Bèle, B., Godoy, E., Dumur, D., 2008b. "Suppression of periodic disturbances in the continuous casting process". In *Proc. Multi-conference on Systems and Control MSC*, San Antonio (US).

COMPUTATIONAL ALGORITHM FOR NONPARAMETRIC MODELLING OF NONLINEARITIES IN HAMMERSTEIN SYSTEMS

Przemysław Śliwiński and Zygmunt Hasiewicz
Institute of Computer Engineering, Control and Robotics, Wrocław University of Technology
W. Wyspiańskiego 27, Wrocław, Poland
przemyslaw.sliwinski@pwr.wroc.pl, zygmunt.hasiewicz@pwr.wroc.pl

Keywords: Hammerstein system, Non-parametric identification, Orthogonal expansions, Regression estimation, Computational algorithm.

Abstract: In the paper a fast computational routines for identification algorithms for recovering nonlinearities in Hammerstein systems based on orthogonal series expansions of functions are proposed. It is ascertained that both, convergence conditions and convergence rates of the computational algorithms are the same as their much less computationaly attractive 'theoretic' counterparts. The generic computational algorithm is derived and illustrated by three examples based on standard orthogonal series on interval, *viz.* Fourier, Legendre, and Haar systems. The exemplary algorithms are presented in a detailed, ready-to-implement, form and examined by means of computer simulations.

1 INTRODUCTION

Recursive routines for nonparametric identification are of interest for practitioners mainly because the *recursive formulas*, involving only the last estimate value and/or the current measurements, are much simpler and much less computationally demanding than their closed-form counterparts, and hence, they seem to be more suitable for applications with limited computational capabilities (*e.g.* in power constrained mobile and/or remote devices).

The advantages of the recursive *orthogonal series* identification algorithms presented here may thus be of importance for a wide range of prospective users, since Hammerstein systems (*i.e.* the cascades of nonlinear static element followed by the linear dynamics; Fig. 1) are a popular modelling tool in many fields, see (Giannakis and Serpedin, 2001); *e.g.* in biocybernetics: (Westwick and Kearney, 2001; Dempsey and Westwick, 2004; Kukreja et al., 2005), chemistry: (Eskinat et al., 1991), control: (Lin, 1994; Zi-Qiang, 1993; Zhu and Seborg, 1994), and in economy: (Capobianco, 2002).

In the paper the new fast routine for a generic orthogonal series algorithm modelling a nonlinear characteristic in Hammerstein systems is proposed and three examples, employing representative orthogonal bases on intervals, are presented. Namely, the following algorithms are provided in a unified and ready-to-implement form:

- the *Fourier* trigonometric,
- the *Legendre* polynomial, and
- the *Haar* wavelet algorithm.

Nonparametric estimates[1] are well known for their flexibility. They allow to model virtually any nonlinearity – be it continuous or not – exploiting the measurement set only, see *e.g.* (Härdle, 1990; Györfi et al., 2002). Application of *orthogonal series*, in particular, enables evaluation of the estimates values in arbitrary points and at any stage of the identification process (in contrast to kernel-based recursive algorithms when the estimation points need to be set beforehand; see *e.g.* (Greblicki and Pawlak, 1989)).

2 REFERENCE ALGORITHM

The Hammerstein system under consideration is described by the discrete-time input-output equation

$$y_k = \sum_{i=0,1,\ldots} \lambda_i m(x_{k-i}) + z_k \tag{1}$$

where $m(x)$ is the system nonlinearity, $\{\lambda_i\}$ is the impulse response of the dynamic subsystem, and z_k is

[1] The term 'nonparametric' refers to the *a priori* knowledge which is at ones disposal rather to the form of the resulting algorithm.

the external, additive noise. The standard nonparametric assumptions are imposed on the system characteristics, input signals and external noise; *cf.* (Greblicki and Pawlak, 1994; Greblicki and Pawlak, 2008; Śliwiński et al., 2009):

1. An input signal, $\{x_k\}$, and an external noise, $\{z_k\}$, are second-order random stationary processes. They are mutually independent and the latter is a zero-mean process. The input $\{x_k\}$ is white and has density, $f(x)$, strictly positive in the identification interval, say a standard unit interval, $[0,1]$.
2. A nonlinear characteristic of the static system, $m(x)$, has ν derivatives.
3. A linear dynamic subsystem is asymptotically stable. Its impulse response, $\{\lambda_i\}$, $i = 0,1,\ldots$, is unknown.
4. A set, $\{(x_l, y_l)\}$, $l = 1,2,\ldots,k,\ldots$ of the system input and output measurements is available.

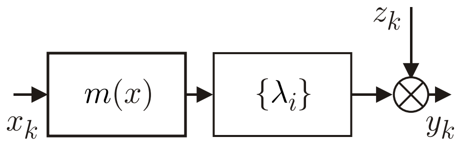


Figure 1: The identified Hammerstein system.

Remark 1. *Due to a composite structure of Hammerstein systems, only a scaled and shifted version of the characteristic $m(x)$ of the static block,* i.e. *the nonlinearity $\mu(x) = am(x) + b$, where $a = \lambda_0 \neq 0$, $b = Em(x_1)\sum_{i=1}^{\infty}\lambda_i$, can at most be recovered from the input-output measurements. Indeed, the following holds,* cf. *(1) and (Greblicki and Pawlak, 2008):*

$$\begin{aligned} E(y_k \,|x_k = x) &= \lambda_0 m(x) + Ez_k \\ &\quad + E\sum_{i=1,\ldots}\lambda_i m(x_{k-i}) \\ &= \lambda_0 m(x) + b \end{aligned}$$

and to recover the genuine $m(x)$ in general case, one needs an additional a priori information about the nonlinearity, e.g. *its value in some points.*

The reference algorithm construction starts with the observation that any square integrable function in the unit interval $[0,1]$ may be represented by the orthogonal series (expansion):

$$\mu(x) = \sum_{m=0}^{\infty}\alpha_m\phi_m(x) \tag{2}$$

where $\{\phi_m\}$, $m = 0,1,\ldots$ is a proper orthonormal basis on the interval $[0,1]$, and where

$$\alpha_m = \langle\phi_m,\mu\rangle = \int_0^1 \phi_m(x)\mu(x)\,dx \tag{3}$$

are the expansion (generalized Fourier) coefficients associated with ϕ_m's. Let $\mu_{\mathfrak{m}}(x)$ be an $\mathfrak{m}$-term approximation (cut-off) of $\mu(x)$, that is, let (*cf.* (2))

$$\mu_{\mathfrak{m}}(x) = \sum_{m=0}^{\mathfrak{m}}\alpha_m\phi_m(x). \tag{4}$$

Due to the completeness of the basis $\{\phi_m\}$ we have

$$\int_0^1 [\mu(x) - \mu_{\mathfrak{m}}(x)]^2\,dx \to 0 \text{ as } \mathfrak{m} \to \infty$$

for virtually any $\mu(x)$; *cf.* (2). Moreover, due to orthogonality of $\{\phi_m\}$, the approximation accuracy grows with the increasing number of approximation terms, $\mathfrak{m}$, as the approximation error is $\sum_{m=\mathfrak{m}+1}^{\infty}\alpha_m^2$.

Assume now that for any k, the earlier and present measurements $\{(x_l, y_l)\}$, $l = 1,\ldots,k$, are sorted (ordered) increasingly with respect to the input values x_l. Then, the orthogonal series reference algorithm may have the following natural form (*cf.* (4))

$$\bar{\mu}_{\mathfrak{m}}(x) = \sum_{m=0}^{\mathfrak{m}}\bar{\alpha}_m\phi_m(x) \tag{5}$$

where (*cf.* (3) and see (Greblicki and Pawlak, 1994; Greblicki and Pawlak, 2008))

$$\bar{\alpha}_m = \sum_{l=1}^{k} y_l \int_{x_{l-1}}^{x_l}\phi_m(u)\,du \tag{6}$$

are estimates of the true expansion coefficients α_m (with $x_0 = 0$). The following theorem describes the limit properties of the reference algorithm:

Theorem 1. *If the number $\mathfrak{m}$ of terms in (5),* i.e. *the number of the estimated coefficients $\bar{\alpha}_m$ in the algorithms, increases with the measurements number k so that*

$$\mathfrak{m} \to \infty \text{ and } \mathfrak{m}/k \to 0 \text{ as } k \to \infty,$$

then

$$E\int_0^1 [\mu(x) - \bar{\mu}_{\mathfrak{m}}(x)]^2\,dx \to 0 \text{ as } k \to \infty.$$

Moreover, for Fourier and Legendre series the algorithm attains, for $\mathfrak{m} = \left\lfloor k^{1/(2\nu+1)} \right\rfloor$, the best possible asymptotic convergence rate, i.e. *for any $\varepsilon > 0$, it holds for them that*

$$E\int_{\varepsilon}^{1-\varepsilon} [\mu(x) - \bar{\mu}_{\mathfrak{m}}(x)]^2\,dx = O\left(k^{-2\nu/(2\nu+1)}\right)$$

while the convergence of the Haar series algorithm achieves, for $\mathfrak{m} = \left\lfloor k^{1/3} \right\rfloor$, the asymptotic rate

$$E\int_0^1 [\mu(x) - \bar{\mu}_{\mathfrak{m}}(x)]^2\,dx = O\left(k^{-2/3}\right)$$

for any $\nu = 1,2,\ldots$.

Proof. The proofs of the theorem for the algorithms with Fourier trigonometric and Legendre polynomial bases can be found in (Greblicki and Pawlak, 1994) and in (Greblicki and Pawlak, 2008). The proof for Haar wavelet algorithm is in (Greblicki and Śliwiński, 2002). □

Remark 2. *Using sorted measurements results in a non-quotient form of the identification algorithms. Such a form (achieved at a moderate cost of keeping the measurement data sorted) can be seen as superior to the alternate quotient-form estimates from the stability and numerical error standpoint, especially, when the number of measurement data is small or moderate (see (Śliwiński, 2009a; Śliwiński, 2009b) for* on-line *and* e.g. *(Greblicki, 1989; Pawlak and Hasiewicz, 1998; Hasiewicz, 1999; Hasiewicz, 2001; Hasiewicz et al., 2005) for off-line quotient orthogonal series algorithms). The orthogonal series algorithm (5)-(6) were presented in (Greblicki and Pawlak, 1994).*

3 COMPUTATIONAL (FAST) ALGORITHM

In a view of Theorem 1, the algorithm (5)-(6) possesses desirable theoretical properties. It however seems not to be computationally attractive for the following two reasons:

- calculating coefficient estimates needs integration, and
- updating the estimates, in case when the new measurement data appear, requires repeating the whole computation routine (6) 'right from scratch'.

Our goal is therefore to make the algorithm computationally efficient without sacrificing its prominent properties. Namely, the abovementioned numeric shortcomings maybe circumvented by:

- avoiding explicit integration in favor of subtraction, and
- providing a computation formula for recursive updating of coefficients estimates.

The goal is accomplished in the following fast generic routine. The first step is elementary – we simply apply here *The First Fundamental Theorem of Calculus* to get the *integration-free* counterpart of the estimate in (6)

$$\bar{\alpha}_m = \sum_{l=1}^{k} y_l \left[\Phi_m(x_l) - \Phi_m(x_{l-1})\right] \tag{7}$$

where $\Phi_m(x)$ are *the indefinite integrals* for the basis functions $\phi_m(x)$. The second step is described in the following proposition (being a generalization of the result presented in (Śliwiński et al., 2009) for wavelets).

Proposition 2. *Let $\bar{\alpha}_m^{(k)}$ denote the estimate of the expansion coefficient α_m obtained for k measurements. Given the ordered sequence, $\{(x_1, y_1), \ldots, (x_l, y_l), (x_{l+1}, y_{l+1}), \ldots, (x_k, y_k)\}$, assume that for the new, $(k+1)$th measurement pair, (x_{k+1}, y_{k+1}), it holds that $x_l < x_{k+1} < x_{l+1}$. Then,* (i) *the new pair is inserted between (x_l, y_l) and (x_{l+1}, y_{l+1}) to maintain the ascending order of the updated measurement set, and* (ii) *the following recurrence formula should be applied to update the coefficient estimates*

$$\begin{aligned}\bar{\alpha}_m^{(k+1)} &= \bar{\alpha}_m^{(k)} + (y_{k+1} - y_{l+1}) \times \\ &\quad \times \left[\Phi_m(x_{k+1}) - \Phi_m(x_l)\right]\end{aligned} \tag{8}$$

with the initial values $\bar{\alpha}_m^{(0)} = 0$, and with the initial measurements set $\{(0,0),(1,0)\}$.

Proof. The proof is immediate. To derive the recurrence formula (8), it suffices to subtract the estimate in (7), computed for *k*, from the one obtained for $k+1$ measurements. □

Below we present three examples showing how to implement Fourier, Legendre, and Haar orthogonal systems in the general identification routine (5)-(8).

3.1 Fourier Trigonometric Series

Since sequence of trigonometric functions

$$\sqrt{1/2\pi}, \left\{\sqrt{1/\pi}\cos(mu), \sqrt{1/\pi}\sin(mu)\right\}$$

constitutes, for $m = 1, 2, \ldots$, an orthogonal basis on the interval $[-\pi, \pi]$; *cf.* (Szego, 1974; Greblicki and Pawlak, 2008), thus for our identification interval, $[0,1]$, we need $\phi_0(x) = 1$ and

$$\begin{aligned}\phi_{2m-1}(x) &= \sqrt{2}\sin((2m-1)\pi x) \\ \phi_{2m}(x) &= \sqrt{2}\cos(2m\pi x)\end{aligned}$$

for $m = 1, 2, \ldots$. From the above we immediately obtain $\Phi_0(x) = x$ and

$$\begin{aligned}\Phi_{2m-1}(x) &= -\tfrac{\kappa}{2m-1}\cos((2m-1)\pi x) \\ \Phi_{2m}(x) &= \tfrac{\kappa}{2m}\sin(2m\pi x)\end{aligned}$$

for $m = 1, 2, \ldots$ and $\kappa = \sqrt{2}/\pi$.

3.2 Legendre Polynomial Series

The Legendre polynomials can be defined recursively as

$$p_{m+1}(x) = \tfrac{2m+1}{m+1} x p_m(x) + \tfrac{m}{m+1} p_{m-1}(x)$$

for $m = 1, 2, \dots$ with $p_0(x) = 1$, $p_1(x) = x$; *cf.* (Szego, 1974; Greblicki and Pawlak, 2008). They form an orthonormal basis on the interval $[-1, 1]$ with the weighting function $\sqrt{(2m+1)/2}$. In our algorithm, for the unit interval we thus need a slightly reformulated

$$\phi_m(x) = \sqrt{2m+1}\, p_m(2x-1).$$

The following recurrence formula for primitives of Legendre polynomials holds (see the derivation in Proposition 3 in Appendix A).

$$P_{m+1}(x) = \kappa_m \left(x^2 - 1\right) p_m(x) + K_m P_{m-1}(x)$$

for $m = 1, 2, \dots,$ where $\kappa_m = (2m+1)/((m+1)(m+2))$, $K_m = m(m-1)/((m+1)(m+2))$ and $P_0(x) = x + 1$, $P_1(x) = \left(x^2 - 1\right)/2$. Eventually, we have

$$\Phi_m(x) = \tfrac{\sqrt{2m+1}}{2} P_m(2x-1).$$

3.3 Haar Wavelet Series

To construct Haar wavelet basis one needs two functions, the father and mother Haar wavelets:

$$\varphi(x) = I_{0 \le x < 1}(x) \text{ and } \psi(x) = \varphi(2x) - \varphi(2x-1)$$

and translations and dilations of the latter, *i.e.*

$$\psi_{kl} = 2^{k/2} \psi\left(2^k x - l\right)$$

where the indices k, l run through the ranges $1, \dots,$ and $0, \dots, 2^k - 1$, respectively; *cf. e.g.* (Wojtaszczyk, 1997). In our case the identification interval, $[0, 1]$, is the native one for Haar system and we can directly take

$$\phi_0(x) = \varphi(x) \text{ and } \phi_m(x) = \psi_{kl}(x)$$

for $m = 1, 2, \dots,$ where

$$k = \lfloor \log_2 m \rfloor \text{ and } l = m \bmod 2^k \qquad (9)$$

and where $x \bmod y = x - y \cdot \lfloor x/y \rfloor$ denotes standard modulus function.

Since, in fact, the father wavelet, $\varphi(x)$, is merely a box function, then the primitives of basis functions, $\phi_m(x)$, are simply

$$\Phi_0(x) = x I_{0 \le x < 1}(x) + I_{x \ge 1}(x)$$

and

$$\begin{aligned} \Phi_m(x) &= \tfrac{1}{\sqrt{2^{k+1}}} \left[\Phi_0\left(2^{k+1} x - l\right) \right. \\ &\quad \left. - \Phi_0\left(2^{k+1} x - (l+1)\right) \right] \end{aligned}$$

for $m = 1, 2, \dots$ with k, l dependent on m and defined as in (9).

4 COMPUTATIONAL COMPLEXITY ANAYSIS

In what follows we compare the computational complexities of both the *reference* and the proposed *fast computational* versions.

4.1 Reference Algorithm

In the reference algorithm implementation one can naturally distinguish two phases with the main routine (5)-(6) preceded by sorting of the measurement sequence. The latter, employing a fast sorting algorithm (*e.g. quick sort*, *heap sort*; *cf.* (Knuth, 1998)), needs $O(k \log k)$ operations.

A naive implementation of the main routine (5)-(6) requires $O(\mathfrak{m}\eta)$ operations in (5) and $O(k\iota)$ operations in (6), where $O(\eta)$ is the cost of evaluating of $\phi_m(x)$ and where $O(\iota)$ is the cost of calculating of the definite integral for $\phi_m(x)$. The overall cost is therefore $O(\mathfrak{m}\eta \cdot k\iota)$. In a view of Theorem 1 this reads $O\left(k^{1+1/(2q+1)} \cdot \eta\iota\right)$. In case of the *Fourier* and *Haar* algorithms we have $O(\eta) = O(1)$. In the *Legendre* algorithm computing $\phi_m(x)$ (*i.e.* a polynomial of order m) takes m operations. The cost $O(\iota)$ of computing integrals (since the indefinite integrals for $\phi_m(x)$ are known) is the same.

Table 1: Complexities of a direct implementation of the reference algorithms.

Algorithm	Cost
Fourier	$O\left(k^{\frac{2}{2q+1}(q+1)}\right)$
Legendre	$O\left(k^{\frac{2}{2q+1}(q+2)}\right)$
Haar	$O\left(k^{\frac{4}{3}}\right)$

4.2 Fast Algorithm

Using the above naive implementation results in cost of at least $O\left(k^{2(q+1)/2q+1}\right)$ operations for every single new measurement to be added. Our algorithm (5), (8) substantially reduces this complexity. First, searching for the pairs (x_l, y_l) and (x_{l+1}, y_{l+1}) in the measurement sequence (employing *e.g.* a standard binary search algorithm) requires $O(\log k)$ operations; *cf.* (Knuth, 1998). Computing the updated value of $\bar{\mu}_{\mathfrak{m}}(x)$ requires another $O(\mathfrak{m}(k))$ operations. The overall cost of the *Fourier* and *Legendre* algorithms (in the latter the recurrence formulas (10), (11) are used) is therefore of order $O(\log k) + O(\mathfrak{m}(k)) = O(\sqrt[2q+1]{k})$.

In case of the *Haar* algorithm, this cost can further be reduced to the order $O(\log k)$ after observation that, due to compactness of Haar functions supports, only $O(\log k)$ terms of are involved in computations of (5), (8). Indeed, using wavelet 'natural' scale-translation notation (*cf.* (9)), one can easily ascertain that, for each scaling index $n = 0, \ldots, \lfloor \log_2 \mathfrak{m}(k) \rfloor$, at most one function $\psi_{nl}(x)$ is non-zero – the one with translation index $l = \lfloor 2^n x \rfloor$. The computation phase of the *Haar* algorithm requires thus only $O(\log k)$ operations for $\mathfrak{m} = \lfloor \sqrt[3]{k} \rfloor$.

Table 2: Complexities of fast implementation of the reference algorithms.

Algorithm	Cost
Fourier	$O\left(k^{\frac{1}{2q+1}}\right)$
Legendre	$O\left(k^{\frac{1}{2q+1}}\right)$
Haar	$O(\log k)$

5 NUMERICAL EXPERIMENTS

The first two examples, the Fourier and Legendre algorithms, possess the same asymptotic behavior while the last, the Haar one, is slightly slower for smooth nonlinearities (*i.e.* for $\nu = 2, 3, \ldots$). However, as we will see in the following numerical experiments, this fact does not necessarily hold true for sample sizes being small or moderate.

To this end, the following piecewise-[smooth|linear|constant] characteristics, referred further to as the *root*, *ramp*, and *step* functions, respectively, were considered in the interval $[0, 1]$:

$$m(x) = \begin{cases} \sqrt[3]{u} \\ 2\left(u+\frac{1}{2}\right) I\left(u+\frac{1}{2}\right) + 2I\left(u-\frac{1}{2}\right) - 1 \\ I(u+1) - I(u) \end{cases}$$

where $u = 2x - 1$ and $I(x)$ is the abbreviated notation of the box function $I_{0 \leq x < 1}(x)$. The number $\mathfrak{m}$ of estimate components, *i.e.* of coefficients in the algorithms, was governed by *the practical rule*, according to which $\mathfrak{m} = \lfloor \sqrt[3]{k} \rfloor$; *cf.* (Greblicki and Pawlak, 2008; Hasiewicz et al., 2005). The input $\{x_l\}$ was uniformly distributed over $[0, 1]$, and the (infinite) impulse response of the dynamic part was $\lambda_i = 2^{-i}$, $i = 0, 1, \ldots$ (thus we had exactly $\mu(x) = m(x)$ for all three nonlinearities, *cf.* Remark 1); the external zero-mean uniform noise was set to make $\max|z_l| / \max|m(x)| = 10\%$. Numerically computed MISE error served as the indicator of algorithms accuracy (computed in slightly narrowed interval, $[\varepsilon, 1-\varepsilon], \varepsilon = 0.1$, in order to avoid the boundary effect affecting Fourier algorithm (*cf.* Figs 2a and 3)).

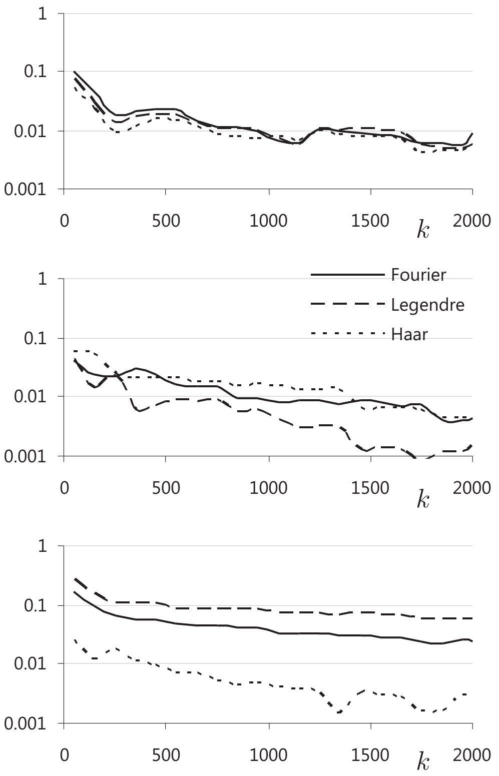


Figure 2: The algorithms errors for three test nonlinearities: **a)** root, **b)** ramp, and **c)** step one.

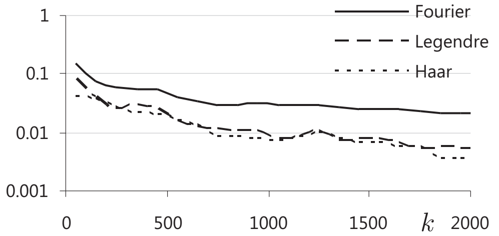


Figure 3: Boundary effect illustration.

The experiments unveil that algorithms offer similar accuracy for the root function. Slightly better performance of Legendre algorithm in case of ramp function and the Haar algorithm in case of step function can both be attributed to similarity of their basis functions to the respective nonlinearities. Nevertheless, the Haar wavelet algorithm – achieving the similar results and being much faster – can be pointed out as the most effective across the whole experiment.

6 FINAL REMARKS

The new class of fast routines for nonparametric identification algorithms recovering the nonlinearity in Hammerstein systems has been proposed. Preserving all the asymptotic properties of their off-line origins, the new algorithms offer much more computationally efficient formulas. Comparing the algorithm properties one can draw the following conclusions:

- *Fourier* algorithm is fast but prone to boundary effect,
- *Legendre* algorithms is the slowest but free boundary problems, finally
- *Haar* algorithm is fast but do not perform well in case of smooth nonlinearities (like the *Fourier* and *Legendre* do).

Remark 3. *Owing to the beneficial features pointed out above it is not a serious disadvantage that all measurement data need to be kept in our algorithm. This – admittedly idiosyncratic feature – is a consequence of both the form of the initial off-line version of the algorithm (6) and the random nature of the input data;* cf. *(Śliwiński et al., 2009). Moreover, the measurement set needs to be maintained only during the synthesis of the estimate. In the implementation step, all k measurements can be rid off and only $\mathfrak{m}$ coefficients (with $\mathfrak{m}$ being a significantly smaller number than k) have to be stored. Observe also that in all nonparametric algorithms, be them kernel or $k-NN$ algorithms, see* e.g. *(Györfi et al., 2002; Greblicki and Pawlak, 2008), the measurements need to be kept as well in order to allow computing the estimate value in arbitrary point.*

That the measurements need to be kept in nonparametric modelling is rather typical as the measurements are essentially the only source of the information about the system/phenomenon. This problem is addressed in (Śliwiński, 2009a; Śliwiński, 2009b) where the quotient form wavelet algorithm is proposed. It is shown there that – on the one hand side – getting rid of the measurements allows the algorithms to be asymptotically equivalent to those possessing all the data, but – on the other – reveals that for small and moderate measurements number such algorithm perform worse.

Finally, we would like to emphasize that the simplicity of the proposed computational algorithm should be seen as an advantage for the practitioners as it allows a straightforward implementation (*cf.* the Appendix).

REFERENCES

Capobianco, E. (2002). Hammerstein system representation of financial volatility processes. *The European Physical Journal B - Condensed Matter*, 27:201–211.

Dempsey, E. and Westwick, D. (2004). Identification of Hammerstein models with cubic spline nonlinearities. *IEEE Transactions on Biomedical Engineering*, 51(2):237–245.

Eskinat, E., Johnson, S. H., and Luyben, W. L. (1991). Use of Hammerstein models in identification of non-linear systems. *American Institute of Chemical Engineers Journal*, 37:255–268.

Giannakis, G. B. and Serpedin, E. (2001). A bibliography on nonlinear system identification. *Signal Processing*, 81:533–580.

Greblicki, W. (1989). Nonparametric orthogonal series identification of Hammerstein systems. *International Journal of Systems Science*, 20:2355–2367.

Greblicki, W. and Pawlak, M. (1989). Recursive nonparametric identification of Hammerstein systems. *Journal of the Franklin Institute*, 326(4):461–481.

Greblicki, W. and Pawlak, M. (1994). Dynamic system identification with order statistics. *IEEE Transactions on Information Theory*, 40:1474–1489.

Greblicki, W. and Pawlak, M. (2008). *Nonparametric system identification*. Cambridge University Press, New York.

Greblicki, W. and Śliwiński, P. (2002). Non-linearity estimation in Hammerstein system based on ordered observations and wavelets. In *Proceedings 8th IEEE International Conference on Methods and Models in Automation and Robotics – MMAR 2002*, pages 451–456, Szczecin. Institute of Control Engineering, Technical University of Szczecin.

Györfi, L., Kohler, M., A. Krzyżak, and Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. Springer-Verlag, New York.

Härdle, W. (1990). *Applied Nonparametric Regression*. Cambridge University Press, Cambridge.

Hasiewicz, Z. (1999). Hammerstein system identification by the Haar multiresolution approximation. *International Journal of Adaptive Control and Signal Processing*, 13(8):697–717.

Hasiewicz, Z. (2001). Non-parametric estimation of nonlinearity in a cascade time series system by multiscale approximation. *Signal Processing*, 81(4):791–807.

Hasiewicz, Z., Pawlak, M., and Śliwiński, P. (2005). Nonparametric identification of non-linearities in block-oriented complex systems by orthogonal wavelets with compact support. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 52(1):427–442.

Knuth, D. E. (1998). *The Art of Computer Programming. Volume 3. Sorting and searching*. Addison-Wesley Longman Publishing Co., Inc, Boston, MA.

Kukreja, S., Kearney, R., and Galiana, H. (2005). A least-squares parameter estimation algorithm for switched Hammerstein systems with applications to the VOR. *IEEE Transactions on Biomedical Engineering*, 52(3):431–444.

Lin, S.-K. (1994). Identification of a class of nonlinear deterministic systems with application to manipulators. *IEEE Transactions on Automatic Control*, 39(9):1886–1893.

Pawlak, M. and Hasiewicz, Z. (1998). Nonlinear system identification by the Haar multiresolution analysis. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, 45(9):945–961.

Śliwiński, P. (2009a). On-line wavelet estimation of nonlinearities in Hammerstein systems. Part I - algorithm and limit properties. *IEEE Transactions on Automatic Control*. Submitted for review.

Śliwiński, P. (2009b). On-line wavelet estimation of nonlinearities in Hammerstein systems. Part II - small sample size properties. *IEEE Transactions on Automatic Control*. Submitted for review.

Śliwiński, P., Rozenblit, J., Marcellin, M. W., and Klempous, R. (2009). Wavelet amendment of polynomial models in nonlinear system identification. *IEEE Transactions on Automatic Control*, 54(4):820–825.

Szego, G. (1974). *Orthogonal Polynomials*. American Mathematical Society, Providence, R.I., 3rd edition.

Westwick, D. T. and Kearney, R. E. (2001). Separable least squares identification of nonlinear Hammerstein models: Application to stretch reflex dynamics. *Annals of Biomedical Engineering*, 29(8):707–718.

Wojtaszczyk, P. (1997). *A Mathematical Introduction to Wavelets*. Cambridge University Press, Cambridge.

Zhu, X. and Seborg, D. E. (1994). Nonlinear predictive control based on Hammerstein system. *Control Theory Applications*, 11:564–575.

Zi-Qiang, L. (1993). Controller design oriented model identification method for Hammerstein system. *Automatica*, 29:767–771.

APPENDIX

Recursion in Legendre Polynomials

The well known recurrence relation between Legendre polynomials of adjacent orders (see *e.g.* (Szego, 1974)), *i.e.*:

$$(m+1)\,p_{m+1}(x) = (2m+1)\,x p_m(x) + m p_{m-1}(x)$$

allows convenient generation of increasing order elements of polynomial orthogonal basis

$$p_{m+1}(x) = \tfrac{2m+1}{m+1} x p_m(x) + \tfrac{m}{m+1} p_{m-1}(x) \qquad (10)$$

for $m = 1, 2, \ldots$, given $p_0(x) = 1$ and $p_1(x) = x$.

In the following proposition we show that the similar relation holds for primitive functions of these polynomials.

Proposition 3. *Let* $P_m(x) = \int_{-1}^{x} p_m(u)\,du$. *The following recurrence relation holds*

$$\begin{aligned} P_{m+1}(x) &= \tfrac{(2m+1)(x^2-1)}{(m+1)(m+2)} p_m(x) \\ &\quad + \tfrac{m(m-1)}{(m+1)(m+2)} P_{m-1}(x) \end{aligned} \qquad (11)$$

for $m = 1, 2, \ldots$ *and with*

$$P_0(x) = x + 1 \text{ and } P_1(x) = \tfrac{1}{2}\left(x^2 - 1\right).$$

Proof. We will give only a sketch of the proof as it involves elementary (yet a bit tedious) calculations. Integrating both sides of the formula in (10) yields

$$P_{m+1}(x) = \tfrac{2m+1}{m+1} \int_{-1}^{x} u p_m(u)\,du - \tfrac{m}{m+1} P_{m-1}(x) \quad (12)$$

Employing now integration by parts and another known recursive formula:

$$\left(1 - x^2\right) p_m'(x) = m\left[p_{m-1}(x) - x p_m(x)\right]$$

we get

$$\int_{-1}^{x} u p_m(u)\,du = \tfrac{x^2-1}{m+2} p_m(x) + \tfrac{m}{m+2} P_{m-1}(x)$$

which applied to (12) yields (11), and (after substitution `m := m - 1`) the formula used in subsequent C++ implementation. □

Code Samples

The following C++ implementations of the presented recursive formulas prove not to be much more intricate than their mathematical origins in (10):

```
template <typename T> struct p
{
 T operator()(T const &x, size_t m)const
 {
    T const _2_ = T(2), _1_ = T(1);
    if(m == 0) return _1_;
    if(m == 1) return  x ;
    p<T> const lp;
    return ((_2_*m-_1_)*x*lp(x, m-1)
              -  (m-_1_)*lp(x, m-2))/m;
 }
};
```

and in (11), respectively:

```
template <typename T> struct P
{
 T operator()(T const &x, size_t m)const
 {
   T const _2_ = T(2), _1_ = T(1);
   if(m == 0) return    x + _1_;
   if(m == 1) return (x*x - _1_)/_2_;
   p<T> const  lp;
   P<T> const lpi;
   return ( (_2_*m-_1_)*(x*x-_1_)
           * lp(x, m - 1)
           + (m - _1_) * (  m - _2_)
           * lpi(x, m - 2))
             /(m * (m + _1_));
 }
};
```

REAL-CODED GENETIC ALGORITHM IDENTIFICATION OF A FLEXIBLE PLATE SYSTEM

S. Md Salleh, M. O. Tokhi and S. F.Toha
Dept.of Automatic Control and System Eng., University of Sheffield, Sheffield, U.K.
cop07sbm@sheffield.ac.uk

Keywords: Real-coded genetic algorithm, Parametric modelling, Flexible plate.

Abstract: Parametric modelling deals with determination of model parameters of a system. Parametric modelling of systems may benefit from advantages of real coded genetic algorithms (RCGAs), as they do not suffer from loss of precision during the processes of encoding and decoding compared with Binary Coded Genetic Algorithm. In this paper, RCGA is used to identify the best model order and associated parameters characterising a thin plate system. The performance of the approach is assessed on basis mean-squared error, time and frequency domain response of the developed model in characterising the system. A comparative assessment of the approach with binary coded GA is also provided. Simulation results signify the advantages of RCGA over two further algorithms in modelling the plate system are also provided.

1 INTRODUCTION

Parametric modelling is defined as the process of estimating parameters of a model characterising a plant. The technique basically searches for numerical values of the parameters so that to give the best agreement between the predicted (model) output and the measured (plant) output. Parametric modelling can include both the parameter estimates and the model structure. Statistical validation procedures, based on correlation analysis, are utilised to validate parametric models.

Several advantages motivating research intention in a flexible structure are due to light weight, lower energy consumption, smaller actuator requirement, low rigidity requirement and less bulky design. These advantages lead to extensive usage of flexible plates in various applications such as space vehicles, automotive industries, and the construction industry. Modelling is the first step in a model-based control development of a system. Accordingly, the accuracy of the model is crucial for the desired performance of the control system.

Artificial intelligence approaches such as genetic algorithm (GA), particle swarm optimisation (PSO), fuzzy logic and neural networks have been utilised in system identification applications. Among these GAs have shown great potential in parametric modelling of dynamic systems.

The utilisation of binary-coded GA (BCGA) and real-coded GA (RCGA) for parameter estimator of models of dynamic systems has been reported in various applications. Zamanan et al. (2006) have reported the use of RGA as an optimization technique for tracking harmonics on power systems. Mitsukura et al. (2002) have reported using BCGA and RCGA to (i) determine a function type and (ii) the coefficient of the function and time delay, respectively. They have tested the technique successfully in determining the hammer stain model and music data model. BCGA also has been used to estimate the parameters of a plate structure (Intan, 2002). However, precision in BCGA is affected due to the processes of encoding and decoding. Moreover, BCGA is susceptible to the Hamming Cliff effect, which can be problematic when searching a continuous search space. Instead of working on the conventional bit by bit operation in BCGA, an RCGA approach is chosen in a wide range of applications where both the crossover and mutation operators are handled with real-valued numbers. A real coded GA leads to reduced computational complexity and faster convergence compared to a binary coded GA.

In this work, RCGA is proposed for parametric modeling of a flexible plate structure in comparison to a binary-coded GA. The rest of the paper is structured as follows: Section 2 describes the flexible plate system and formulates the problem. Section 3 presents the parametric models with RCGA and parametric system identification respectively. Section 4 presents implementation of the algorithms in modeling the system using various excitation signals such as finite duration step, random and pseudo random binary signal (PRBS). Results and discussions of the model validity through input/output mapping, mean square of output error and frequency domain response are also presented. Parametric modelling is also confirmed with convergence of fitness values and time run. Finally, the paper is concluded in Section 5.

2 THE FLEXIBLE PLATE SYSTEM

Dynamic simulation of a plate structure using the finite differences (FD) method is considered in this paper. The finite difference method is used to discretise the governing dynamic equation considered with no damping and the lateral deflection of plates is obtained using central finite difference method. It then transformed into state space equation as the following equation.

$$\boldsymbol{W}_{i,j,k+1} = (\mathbf{A}+2_{ijk})\boldsymbol{W}_{i,j,k} + \boldsymbol{BW}_{ijk} + \boldsymbol{C}\boldsymbol{F} \qquad (1)$$

Where 2_{ijk} represents the diagonal elements of *(2/c), $C=(\Delta t^2/\rho)$, $c=-DC$*, and $\boldsymbol{W}_{i,j,k+1}$ is the deflection of grid points $i = 1, 2, \ldots\ldots, n+1$ and $j = 1, 2, .., m+1$ at time step k+1. $\boldsymbol{W}_{i,j,k}$ and $\boldsymbol{W}_{ijk}$ are the corresponding deflections at time steps k and k-1 respectively. **A** is constant $(n+1)(m+1)$ x $(n+1)(m+1)$ matrix whose entries depend on physical dimensions and characteristics of the plate, $\boldsymbol{B}$ is a diagonal matrix of -1 corresponding to $\boldsymbol{W}_{i,j,k}$ and C is a scalar related to the given input and $\boldsymbol{F}$ is an $(n+1)(m+1)$ x 1 matrix known as the forcing matrix. The algorithm is implemented in Matlab/SIMULINK with applied external force or disturbance into all clamped edges plate. Twenty two equal divisions of plate elements with dimension 1.0mm× 1.0mm× 0.00032m is measured at the detection and observation points (Figure 1). Parameters of the plate considered comprise mass density per area, ρ = 2700 kg/m^2, Young's Modulus, E = 7.11 x 10^{10} N/m^2, second moment of inertia, I = 5.1924 x 10^{-11} m^2 and Poisson ratio, υ = 0.3 with sampling time 0.001.

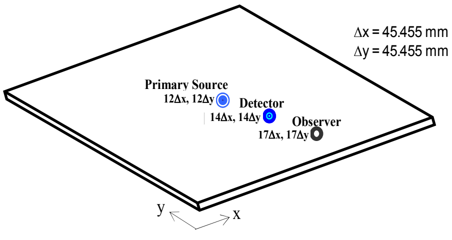


Figure 1: The flexible plate system.

3 REAL CODED GENETIC ALGORITHM

In most of practical engineering problems, the real-coded GA is more suitable than the binary-coded GA, as transformations from real number to binary digits may suffer from loss of precision. Genetic operations are very important to the success of specific GA applications. In this work, real-coded representation is used to determine the model order of the plant and subsequently identify parametric model of the system. The initial population is created randomly within [-1,1] range. The main three genetic operators involved are described below.

3.1 Selection

Selection is the process of determining the number of times or trials a particular individual in the population is chosen for reproduction (Chipperfield, 1994). The process includes two steps, namely selection probability and sampling algorithm. Selection probability is concerned with transformation of raw fitness values into real as expected of an individual to reproduce. Sampling algorithm reproduces individuals based on the selection probabilities computed before. This process is repeated as often as individuals must be chosen. There are many methods reported such as roulette wheel selection, stochastic universal sampling and tournament selection, etc. The stochastic universal sampling (SUS) method is used in this work that randomly copies chromosomes and simulates N equally distributed pointers. SUS is a simpler algorithm, and as individuals are selected entirely on their position in the population, SUS has zero bias. After selection has been carried out, the construction of the intermediate population is complete and the crossover and mutation operators are then applied.

3.2 Crossover (Recombination)

Crossover produces new individuals that have some parts of both parent's genetic material (Chipperfield, 1994). However, Mühlenbein et. al (1991) have distinguished between recombination and crossover. The mixing of the variables was called recombination and the mixing of the values of a variable was named crossover. Line recombination employed in this work performs an exchange of variable values between the individuals. By using a real-valued encoding of the chromosome structure, line recombination is a method of producing new phenotypes around and between the values of the parents' phenotypes (Mühlenbein and Schlierkamp, 1993). For the line recombination, let $x = (x_1,...,x_n)$ and $y = (y_1,...,y_n)$ be the parent strings. Then, the offspring $z = (z_1,...,z_n)$ is computed by

$$z_i = x_i + \alpha(y_i - x_i) \qquad i = 1,...,n \tag{2}$$

where α is chosen uniform randomly in [-0.25, 1.25]. Each variable in the offspring is the result of combining the variables in the parents according to (2). Line recombination can generate any point on the line defined by the parents within the limit of the perturbation, α, for a recombination in two variables. This operator can overcome limitations in variables decision and help improve in exploration during recombination.

3.3 Mutation

The mutation operator arbitrarily alters one or more components, *genes*, of a selected chromosome so as to increase the structural variability of the population. The role of mutation in GAs is that of restoring lost or unexplored genetic material into the population to prevent the premature convergence of GA to suboptimal solutions; it insures that the probability of reaching any point in the search space is never zero. Each position of every chromosome in the population undergoes a random change according to a probability defined by a mutation rate, the mutation probability, p_m (Herrera et.al, 1998). The probability of mutating a variable is set to be inversely proportional to the number of variables (dimensions). The more dimensions one individual has the smaller the mutation probability of it will be. A mutation rate of 1/*m,* (where *m* is the number of variables) produced good results for a broad class of test function. However, the mutation rate was independent of the size of the population (Mühlenbein and Schlierkamp, 1993). The mutation operator for the real coded GA uses a non-linear term for the distribution of the range of mutation applied to gene values. Real value mutation is used in this work.

3.4 The Fitness Function

In this study, minimum mean square error is used as a fitness function of the algorithm, while number of generations is used as stopping criterion. The fitness function, X, is set to minimize (3), in such a way that it approaches zero;

$$X = \min \left(\frac{1}{n} \sum_{i=1}^{n} \left(\left| y(i) - \hat{y}(i) \right| \right)^{2} \right) \tag{3}$$

where *y(i)* is the actual system output subjected to a disturbance signal, $\hat{y}(i)$ is the response of the estimated system under the same disturbance, and i=1,2,…,n ; n is total number of input/output sample pairs. The algorithm of all executions predefined a maximum number of generations as stopping criteria.

3.5 Values of Real-coded Genetic Parameters

The real-coded GA parameters used are presented in Table 1.

Table 1: Parameters of real-coded GA.

RCGA Properties	
Population Size	100
Selection rate	0.9
$P_{c,max}$, $P_{c,min}$	0.67
$P_{m,max}$, $P_{m,min}$	1/n (n=no of variables)
Selection Method	SUS
Crossover Method	Line Recombination
Mutation method	Real-value mutation

4 PARAMETRIC SYSTEM IDENTIFICATION

The transfer function of the model used corresponds to the ARMA model structure by neglecting the noise, η term;

$$\begin{aligned} \hat{y}(k) &= -a_1 y(k-1) - \ldots - a_4 y(k-4) \\ &+ b_0 u(k-1) + \ldots + b_3 u(k-4) \end{aligned} \tag{4}$$

In matrix form, the above equation can be written as

$$\begin{aligned} \hat{y}(k) &= -\begin{bmatrix} a_1, a_2, \\ a_3, a_4 \end{bmatrix} \begin{bmatrix} y(k-1), y(k-2), \\ y(k-3), y(k-4) \end{bmatrix}^{T} \\ &+ \begin{bmatrix} b_0, b_1, \\ b_2, b_3 \end{bmatrix} \begin{bmatrix} u(k-1), u(k-2), \\ u(k-3), u(k-4) \end{bmatrix}^{T} \end{aligned} \tag{5}$$

The first four variables are assigned to $b_0,...,b_3$ and the next four to $a_1,...,a_4$ as indicated in (5). Once the model is determined, the model needs to be verified to determine whether it is well enough to represent the system. Correlation tests including autocorrelation of the error, cross correlation of input-error, input*input-error are carried out to test and validate the model. Each simulation was observed over 7000 samples of data for each set. The first five resonance frequencies of vibration of the plate found from spectral density of the predicted output of the RCGA model were 9.971 rad/s, 34.51 rad/s, 56.76 rad/s, 78.23 rad/s and 99.71 rad/s.

5 RESULTS

In order to determine appropriate model order for system model using RCGA, different model orders were tested. The results of these tests with model orders of 4 to 12 are summarized in Table 2. The results include time run, standard deviation, mean value and mean square error. The accuracy of the model, for different model orders, is presented in terms of standard deviation, mean value and MSE normalized with 10^{-15}, run time represented in minutes, and values averaged for each 5 runs. As noted in Table 2, a model order of 4 achieved minimum mean square error of 1.195 with the smallest standard deviation computational time, and this was thus chosen for obtaining a model of the flexible plate.

Table 2: Accuracy of model order.

Model Order	4	6	8	10	12
Std. Deviation	5.825	6.492	10.10	7.332	10.86
Mean Value	2.208	2.383	2.996	3.097	3.939
Normal MSE	1.195	1.203	1.196	1.281	1.562
Time Run (min)	34.84	34.93	42.55	42.13	43.02

In subsequent attempts, model order of four (4) has been used to obtain unknown parameters of RCGA model system in comparison to binary coded genetic algorithm (BCGA). In BCGA, the design parameters are similar to those in RCGA with single point crossover and mutation rate of 0.0001. For RCGA, the time-domain and frequeny-domain results with random disturbance are shown in Figure 2 and Figure 3 respectively. Both figures show agreement between the actual and predicted output in modelling the plate. The normalized error between the two outputs as depicted in Figure 4 is reasonably small. The corresponding correlation test results are shown in Figure 5 using random signals for RCGA, and these are in general within the 95% confidence level. Thus, this confirms the accuracy of the model in representing the dynamic behaviour of the plant system.

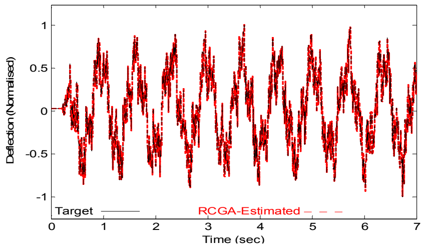


Figure 2: The error between actual- predicted outputs.

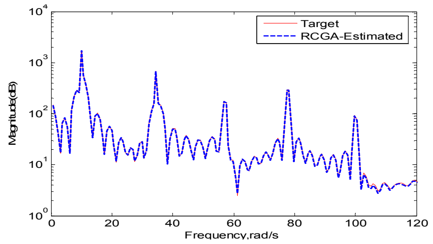


Figure 3: PSD of the actual-predicted outputs.

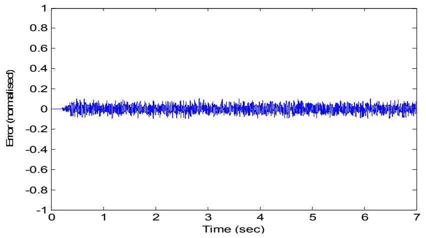


Figure 4: Error between the actual-predicted outputs.

Small or less significant parameter variations with BCGA indicate convergence to local minima and/or pre-mature convergence. The MSE values achieved after 500/1000 generations (Figure 6 – Figure 8) with BCGA and RCGA are shown in Table 3. RCGA achieved faster convergence compared to BCGA. The RCGA achieved better convergence than BCGA over 500 generations or less

Table 3: Mean squared output error with the Gas.

Algorithm (Generation)/ Disturbance	Mean Squared Error		
	Random ($x10^{-4}$)	PRBS ($x10^{-4}$)	Step ($x10^{-6}$)
RCGA (500)	9.51350	1.83940	1.022
RCGA (1000)	9.51070	1.84130	1.199
BCGA (500)	12.02200	4.41720	7.6564

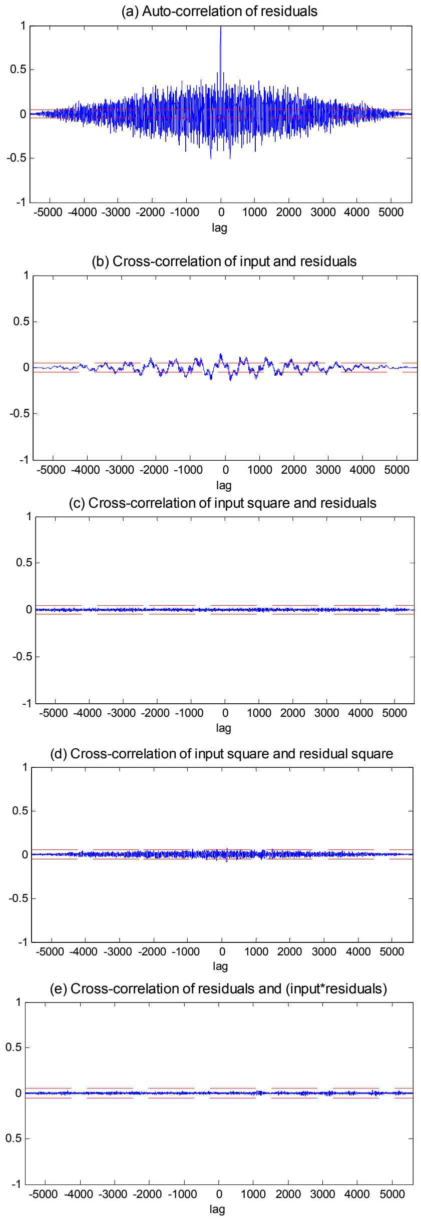


Figure 5: Correlation validation tests (a) – (e).

(recommended about 350) with all the test signals. It was noted that a larger number of generations did not improved the convergence rate, but took more time to compute. Figures 9 and 10 show the convergence of parameter estimates with RCGA as compared to BCGA.

The estimated system model parameters [a1, a2, a3, a4, b0, b1, b2, b3] with the tested disturbance signals at the end of 500 generations with RCGA and BCGA are shown below.

i) Random disturbance

RCGA: [0.07336, 0.1579, 0.1716, 0.07099, 1, 0.5824, –1, 0.392],

BCGA: [–0.375, 0.6445, –0.107, 0.1354, 0.9176, 0.5837, –0.7937, 0.2734]

ii) PRBS

RCGA: [0.1355, –0.2193, 0.3892, –0.2897, 1,0.6084, –1, 0.3739]

BCGA: [–0.5263, 0.3177, 0.0453, 0.3203, 1, 0.3285, –0.9275, 0.5801]

iii) Finite duration step

RCGA: [0.1850, –0.0002, –0.5244, 0.3418, 1, 0.4964, –0.0352, –0.4639],

BCGA: [–0.0576, 0.7715, –0.9993, 0.3186, 0.4695, 0.3206, 0.4653, –0.4607]

Figure 11 shows the MSE (in 10^{-4}) and associated computer run time (in hours) for convergence with RCGA and BCGA. It is noted that in general the RCGA required less computing time as well as achieved lower MSE values as compared to BCGA.

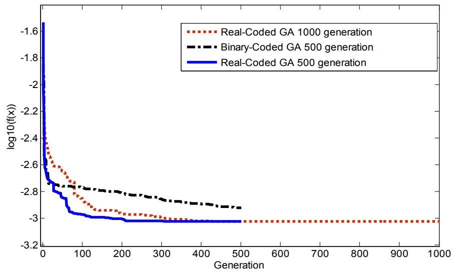


Figure 6: Convergence with random signal.

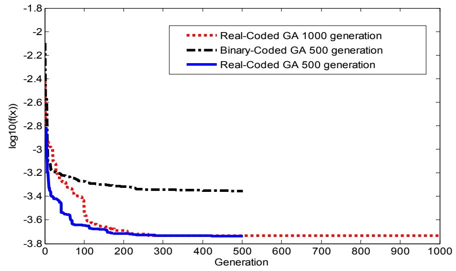


Figure 7: Convergence with PRBS Signal.

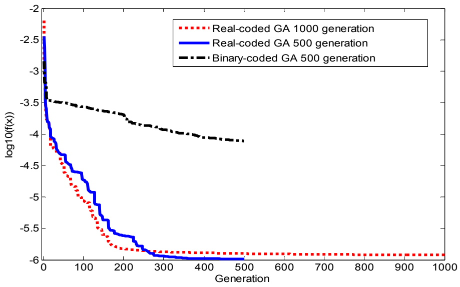


Figure 8: Convergence with step signal.

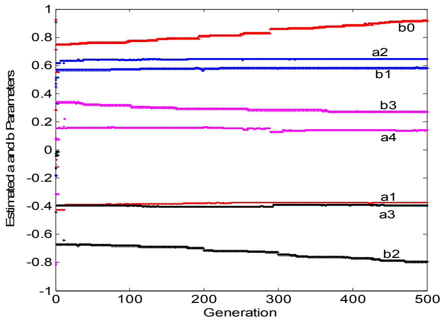


Figure 9: Estimated parameters with BCGA.

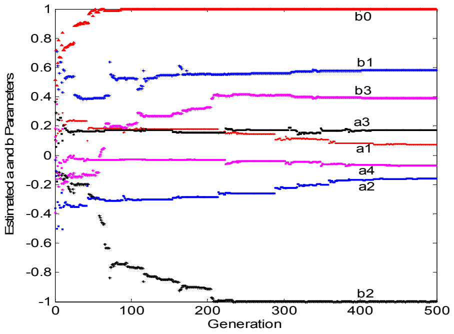


Figure 10: Estimated parameters with RCGA.

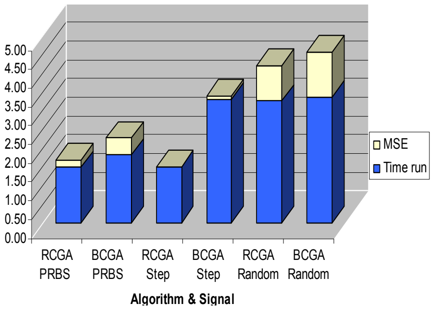


Figure 11: MSE and time run for GAs.

6 CONCLUSIONS

Parametric modelling of a flexible plate system has been carried out. Real-coded GA has been used for estimation of order and parameters of the model characterising the dynamic behaviour of the plate system. The approach has been evaluated in comparison to equivalent binary-coded GAs with three different test signals. It is noted that the models obtained with RCGA have performed better in characterising the system in comparison to those obtained with BCGA.

ACKNOWLEDGEMENTS

The author acknowledges the support of a research fellowship of the University Tun Hussein Onn Malaysia (UTHM) and Ministry of Higher Education Malaysia.

REFERENCES

Chipperfield, A. J., Fleming, P. J., Pohlheim, H. and Fonseca, C., 1994. A genetic algorithm toolbox for MATLAB, *Proceedings of the International Conference on Systems Engineering*, Coventry, UK.

Herrera, F., Lozano, M., Verdegay, J.L., 1998. Tackling real-coded genetic algorithms: operators and tools for behavioural analysis, *Artificial Intelligence Review*, Vol. 12, pp. 265-319.

Mat Darus, I.Z., 2004. *Soft computing Adaptive Active Vibration Control of Flexible Structures*, PhD Thesis, Dept. of Automatic Control and Systems Engineering, The University of Sheffield, Sheffield, UK.

Md Salleh, S., Tokhi, M.O., “Discrete Simulation of a Flexible Plate Structure using a State-Space Formulation”, *Proceedings of 7th International Conference on System Simulation and Scientific Computing (ICSC’2008)*, Beijing China, 10-12 Oct 2008.

Mitsukura, Y.,M., Fukumi, Norio Akamatsu and Yamamoto,T. , 2002. A System Identification Method Using a Hybrid-Type Genetic Algorithm, *Proceedings of the 41st SICE Annual Conference*, Vol.3 ,pp.1598-1602.

Mühlenbein, H., and Schlierkamp-Voosen, D., 1993. Predictive Models for the Breeder Genetic Algorithm: I. Continuous Parameter Optimization, *Evolutionary Computation,* Vol.1, issue 1, pp. 25-49.

Mühlenbein, H. , Schomisch, M., and Born, J., 1991. The parallel genetic algorithm as function optimizer, *Parallel Computing*, 17, pp. 619-632.

Shaheed, M.H. and Tokhi, M.O., 2002. Dynamic modelling of single-link flexible manipulator: parametric and non-parametric approaches, *Robotica*, **20**, pp. 93–109.

Zamanan, N.; Sykulski, J.K.; Al-Othman, A.K., Real Coded Genetic Algorithm Compared to the Classical Method of Fast Fourier Transform in Harmonics Analysis, *Proceedings of the 41st International UPEC '06*, Vol. 3, pp. 1021-1025.

MODEL-BASED DESIGN OF CODE FOR PLC CONTROLLERS

Krzysztof Sacha
Warsaw University of Technology, Nowowiejska 15/19, 00-665 Warszawa, Poland
k.sacha@ia.pw.edu.pl

Keywords: Automatic program generation, Model verification, Finite state machine, Programmable logic controller.

Abstract: This paper describes a method for model-based development of software for programmable logic controllers (PLC). The method includes modeling of a control algorithm, verifying the algorithm with respect to the requirements, and automatically generating the code in one of the IEC 61131 languages. The modeling language is UML state machine diagram, and the verification tool is UPPAAL model-checking toolbox. The method has good scalability with respect to the number of the modeled objects and the ability to cope with integer values by means of variables and function blocks.

1 INTRODUCTION

This paper describes a method for model-based development of software for programmable logic controllers – PLC. The method includes modeling of a control algorithm, verifying the algorithm, and automatically generating the code for a PLC.

The development cycle is shown in Figure 1. The modeling language is UML state machine diagram (OMG, 2005), which has been widely accepted as a means for specifying the controller at a suitable high level of abstraction. The verification tool is the UPPAAL model-checker (Behrmann et al, 2004). When the verification has been finished, the implementation code can be generated automatically in one of the IEC 61131 languages (IEC, 1993).

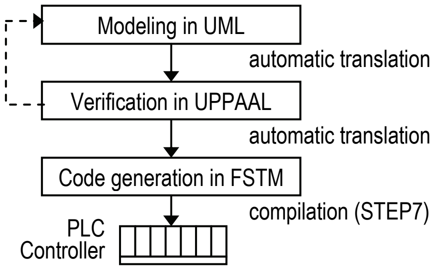


Figure 1: Modeling, verification and implementation of the program code.

A formal semantics for a UML state machine is given by a translatable finite state time machine – FSTM (Sacha, 2007, 2008). Modeling a controller in UML, modeling the environment in UPPAAL, and formulating safety requirements in a formal language of CTL formulae are done manually. The tasks of converting the model from UML to UPPAAL and to FSTM, verifying the model, and generating the program code are done automatically.

The unique features of the method described in this paper are the use of UML state machine as a problem modeling tool, and the ability to verify time dependent behavior of the controller. Widely accepted models of timed automata (Alur, Dill, 1996) and timed I/O automata (Kaynar et al, 2006) are used mainly for modeling and verification of time-dependent behavior of state systems. Still another models of time triggered automata (Krcal et al, 2004) and PLC-automata (Dierks, 1997) are used for code generation only.

The paper is organized as follows. Section 2 gives an overview of PLC controller and finite state time machine. Section 3 defines the semantics of UML state machine in FSTM. Section 4 presents a conversion algorithm from FSTM to UPPAAL and explains the verification process. The conversion of finite state time machine into a program code is described in Section 5. A discussion of the results and plans for future work are given in Conclusions.

2 PLC CONTROLLER

PLC is a computerized device that cooperates with its environment through a set of input and output signals. The controller executes in a loop, polling the inputs and computing the values of the outputs.

The controller counts time using timers. A timer can be activated in a given a set of states. An active timer counts time and expires when it has continued to be active for a predefined period of time. An expired timer is perceived by the controller similarly as an input signal. The execution of a controller can be described in a pseudo-code, which creates a reference model for PLC execution:

```
state  = initial_state();
loop_forever {
   input  = poll_the_input();
   timers =
   set_timers(active_timers(state));
   state  = next_state(state,timers,input);
   output = count_output(state);
   set_the_output(output);
}
```

Board software of a PLC sets the initial state (`initial_state`), executes the loop (`loop_forever`), polls the input signals (`poll_the_input`) counts time and sets the expired timers (`set_timers`), and sets the output signals (`set_the_output`). The programmer must only write a code for selecting active timers (`active_timers`), and calculating the next state of the controller (`next_state`) and of the output (`count_output`).

The semantics of a PLC program is defined by a finite state time machine (Sacha, 2007), which is a tuple $A = (S, \Sigma, \Gamma, \tau, \delta, s_0, \Omega, \omega)$ where

S is a finite set of *states*,
Σ is a finite set of *input symbols*,
Ω is a finite set of *output symbols*,
Γ is a finite set of variables called *timer symbols*,
$\tau: \Gamma \rightarrow 2^S \times N^+$ is an injective function, called *timer function* (with two projections $\tau_S: \Gamma \rightarrow 2^S$ and $\tau_N: \Gamma \rightarrow N^+$, respectively),
$\delta: S \times \Sigma \times 2^\Gamma \rightarrow S$ is a partial function, called *transition function*, such that:
$[(s, a, T) \in Dom(\delta)] \Leftrightarrow (\forall t \in T)[s \in \tau_S(t)]$
$s_0 \in S$ is the initial state,
$\omega: S \rightarrow \Omega$ is an output function.

Notation: N^+ is the set of positive integers, $Dom(\delta)$ is the domain of a function δ, $card(X)$ is the cardinality of a set X, and ϕ is an empty set.

Finite state time machine looks much like a Moore automaton with three additional elements: Γ, τ, ε, which add to the model the dimension of time. A timer symbol $t \in \Gamma$ is a variable, which takes values from the set N^+. The current value of t is interpreted as the duration of a period of time. Timer function τ assigns to each timer a group of states and a constant value. The meaning is such that timer t is enabled, i.e. counts time, as long as the automaton resides in one of the states from $\tau_S(t)$ and it expires when the current value of t exceeds $\tau_N(t)$.

Timer symbols in Γ can be set in an arbitrary order and denoted $t^1 \dots t^n$. The valuation t of timer symbols can be described as a vector of values t. The current value of a timer t^i is denoted t^i.

The execution of a finite state time machine starts in state s_0 with the values of all timers equal to 0. For a given state s_k and a valuation of timers t_k there exists a set of expired timers, defined as:

$$\Theta(s_k, t_k) = \{ t^i \in \Gamma : s_k \in \tau_S(t^i) \text{ and } t^i_k \geq \tau_N(t^i) \}$$

The machine executes in a state (s_k, t_k) by taking an input symbol a_k and moving to the next state s_{k+1} defined by the transition function:

$$s_{k+1} = \delta(s_k, a_k, \Theta(s_k, t_k)) \qquad where\ k = 0, 1, ...$$

When the machine enters a state s_{k+1} time advances and the values of timers change reflecting the elapsed time interval:

$$t^i_{k+1} = \begin{cases} t^i_k + 1 & \text{if } s_{k+1} \in \tau_S(t^i) \text{ and } s_k \in \tau_S(t^i) \\ 0 & \text{otherwise} \end{cases}$$

When the valuation of timers t changes, the set Θ of expired timers may change as well. This way a finite state time machine can respond to the flow of time, even if $s_{k+1} = s_k$ and $a_{k+1} = a_k$. Please note that the last argument of δ is a set of expired timers, hence, no conflict exists if several timers expire at the same time instant.

The state space of a PLC as well as of an FSTM can be defined by enumerating all of the elements, eg. $S = \{s_1, s_2, ..., s_n\}$. An alternative way is to allow for using variables and to define the state space as a Cartesian product of a set of enumerated elements and a set of all possible valuations of those variables. This is only a shorthand notation, which does not add any new semantics to the model, and therefore it is not shown in the formal definition.

In the rest of this paper, we will adopt a naming convention of UPPAAL (Behrmann et al, 2004) and refer to the enumerated elements of state as locations. Locations will be shown in graphical models explicitly, as the nodes of a graph, while variables will be referred to by guard expressions and will be assigned values within actions of transitions.

3 UML STATE MACHINE

UML state machine diagram is a graph composed of nodes, which are locations, and edges, which are labeled transitions. A transition can be triggered by a signal received from the outside. A transition which is triggered can fire, if the corresponding guard expression over a set of variables evaluates to true. Firing of a transition can move the machine to a new location, change the values of variables and send a signal. This way, the state space of a UML state machine is a Cartesian product of the set of locations and the set of all possible valuations of variables.

UML allows for nesting of locations. However, a hierarchy of locations can always be flattened. A formal model and an algorithm for flattening the hierarchy were described in detail in (Sacha, 2007) an will not be discussed in the rest of this paper.

Relating this model to a PLC, one can note that a received signal corresponds to a combination of the input signals of the PLC, and a sent signal corresponds to a combination of the output signals. States of an UML state machine and transitions between states correspond to states of a PLC and to the next-state function defined by a program code.

A conversion algorithm of a UML state machine into a FSTM can be described as follows.

S equals to the Cartesian product of the set of all locations of the UML state machine and the valuations of variables used in guard expressions.

Σ equals to the set of external signals, which trigger transitions in the UML state machine; a signal is a combination of all the input signals of the PLC.

Γ is a set of timer symbols $t^1,...,t^n$; there is one timer symbol t^i for each timed transition (i.e. transition with an *after* clause) in the UML state machine,

τ is the timer function, which assigns to each timer symbol t^i created for a timed transition T a pair composed of a source state of this transition and the value of the after clause of this transition.

δ is the transition function $\delta: S \times \Sigma \times 2^{\Gamma} \rightarrow S$, such that: $\delta\,(s_1\,, a, T) = s_2$ if and only if there exists a transition in the UML state machine diagram such that s_1 is the source and s_2 the destination state of this transition, and either a is the event that triggers this transition (in this case $T = \phi$), or $T = \{t^i\}$ and t^i is the timer symbol of this timed transition (ie. $\delta\,(s_1, a, T) = s_2$ for all $a \in \Sigma$).

s_o is the initial state of the UML state machine.

Ω equals to the set of combinations of all the output signals of the PLC that are set by the actions of the UML state machine.

ω is the output function, which assigns to each state $s \in S$ the output symbol $q \in \Omega$, which is set by all transitions to s.

Example. Consider a railroad crossing controlled by a PLC. There are a number of railway tracks within the crossing, and a number of trains can approach the crossing simultaneously (one train on a track is allowed). The movement of trains is controlled by a set of semaphores that can prevent trains from entering the crossing. The road traffic is controlled by a gate that can be *open* or *closed*. A semaphore can be operated by a controller to display *green* light, when a train approaches, but not earlier than after the gate has been closed. Opening and closing states of the gate are confirmed to the controller by the input signals: *up* and *down*, respectively. Closing the gate must last less than 30 seconds, or else an alarm must sound. The semaphores are *red* and the gate is *up* in the initial state of the crossing.

An algorithm for the railroad crossing controller is shown in Figure 2. The locations within the graph correspond to states of the crossing with respect to train positions. The transitions bear labels of the type *event / action*, where *event* corresponds to a condition on the input signals or timers, and *action* corresponds to setting the values of variables.

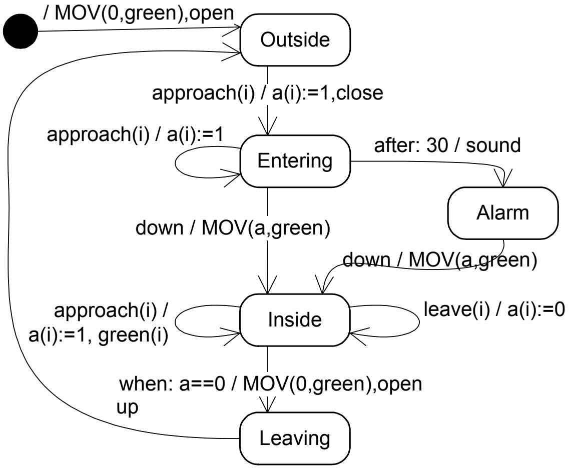


Figure 2: UML model of the railroad crossing controller.

The positions of particular trains are signaled to the controller by short input pulses *approach(i)* and *leave(i), i=0,...n-1*. The appearance of an *approach*-pulse is stored in a vector variable *a(i), i=0,...n-1* and makes the controller to close the *gate*. When the gate is *down*, the controller uses the stored data to send *green* signals to the appropriate semaphores – the function *MOV(a,green)* sends a *green*-signal for each train, which approaches the crossing.

The controller keeps track of all the trains inside the crossing, and waits until the last train has left. If this is the case, the controller turns the *green* signals off, *opens* the gate and waits until the gate is *up*.

Vector *a* is part of the controller state. This way, there are in fact as many *Entering* and *Inside* states as are the combinations of values in vector *a*. Output signals of the state machine are *open* and *close* to operate the gate, and the signals *green(i)* to operate the semaphores to display *green* or *red*.

FSTM model of the controller has the same set of locations and the same set of variables. It has a single timer symbol *t*, and the timer function $\tau_S(t) = \{Entering\}$ and $\tau_N(t) = 30$. The transition function is defined by the set of all the transitions of the UML state machine. The sets of input and output symbols are the combinations of input and output signals.

4 VERIFICATION

UPPAAL is a toolbox for modeling and verification of real time systems, based on the theory of timed automata. The core part of the toolbox is a model-checking engine, which enables for verification of properties defined as CTL path formulae.

A timed automaton (Alur, Dill, 1996), as used in UPPAAL, is a finite state machine extended with clock variables that evaluate to positive real numbers and state variables that evaluate to discrete values. State variables are part of the state. All the clock variables progress simultaneously. An automaton may fire a transition in response to an action, which can be thought of as an input symbol, or to a time action related to the expiration of a clock condition. A clock variable can be reset to zero at a transition.

A set of timed automata can be composed into a network over a common sets of clocks, variables and actions. This way a cooperation between a controller and a controlled plant can be modeled.

The use of a dense-time model-checker to verify a discrete-time model may look as an overkill. But in fact it is not, because the environment of the controller works in real-time and must be modeled using a dense-time method.

A conversion algorithm of FSTM into UPPAAL is described in (Sacha, 2008).

Verification. UPPAAL can verify the model with respect to the requirements, expressed formally as CTL formulae. To do this, UPPAAL model-checker evaluates path formulae over the reachability graph of a network of timed automata.

The query language consists of state formulae and path formulae. A state formula is an expression that can be evaluated for a particular state in order to check a property (e.g. a deadlock). Path formulae quantify over paths of execution and ask whether a given state formula φ can be satisfied in any or all the states along any or all the paths.

Path formulae can be classified into three types:

- Reachability properties: $E<>\varphi$. (will φ be satisfied in a state of a path?)
- Safety properties: $E[]\varphi$ and $A[]\varphi$. (will φ be satisfied in all the states along a single or along all paths?)
- Liveness properties: A<>φ and ψ-->φ. (will φ eventually be satisfied? will φ respond to ψ?)

Example. Consider again the railroad crossing described in Section 3. A train cannot be stopped instantly. When a train is detected by a train position sensor, a controller has 30 seconds to *close* the gate and display a *green* signal, which allows the train to continue its course. After these 30 seconds, it takes further 20 seconds to reach the crossing. Otherwise, if the *green* signal is not displayed within these 30 seconds, the train must break in order to stop safely before the crossing. Closing the gate must last less than 20 seconds, or else an alarm must sound. The gate can be opened when the position sensor has sent a *leave* signal after the last train has left the crossing.

The environment of the controller consists of a number of trains and a gate. Each of these elements can be modeled in UPPAAL and synchronized with the controller within a network of timed automata.

The template of a train is shown in Figure 3. Actions, which names bear the suffix '?', act like input symbols that enable the associated transitions. Actions, which names bear the suffix '!', act like output symbols that are passed to other automata in order to trigger the respective input symbols. This way the execution of one automaton can control the execution of a other automata.

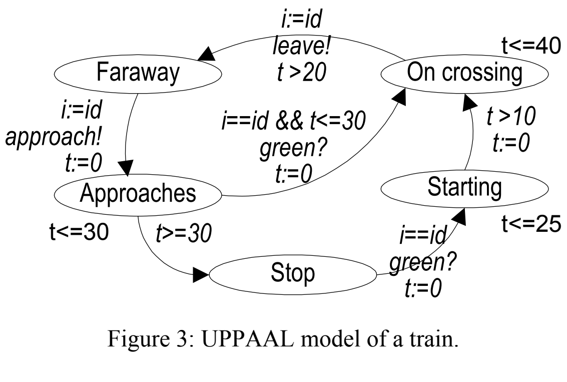


Figure 3: UPPAAL model of a train.

Time invariant *t≤30* of state *Approaches* forces a transition after 30 seconds have passed since the train has entered the state. This models the necessity of breaking the train if *green* has not been displayed in time. Time condition *t>20* at the transition from *On crossing* to *Faraway* reflects the minimum time of passing the crossing by a fast train. Time invariant *t≤40* of the state *On crossing* reflects the maximum time of passing by a slow train.

The template is parameterized with the train identifier *id*. A set of *n* trains, e.g. four, can be generated using the values of *id =0* through *3*.

A model of the gate is shown in Figure 4. Time invariants *t ≤20* at states *Closing* and *Opening* reflect time that it takes to close or to open the gate.

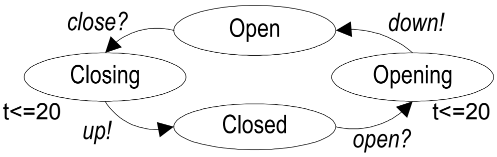


Figure 4: UPPAAL model of the gate.

The simple reachability properties can check if a given state is reachable:

- *E<> train1.On crossing*: This checks if train 1 can pass the crossing (a similar property can be checked for other trains).
- *E<> (train1.On crossing && train2.On crossing && train3.On crossing && train4.On crossing)*: This checks if all the trains can move through the crossing simultaneously.

The safety properties can check that unsafe states will never happen:

- *A [] (train1.On crossing or train2.On crossing or train3.On crossing or train4.On crossing) imply gate.Closed*: This ensures that each time a train is passing the crossing, the gate is closed.
- *A [] (gate.open imply (¬ train1.On crossing && ¬ train2.On crossing && ¬ train3.On crossing && ¬ train4.On crossing))*: This ensures that each time the gate is open, a train is not on the crossing.

The liveness properties can check consequences of an event, e.g.:

- *train1.Approaches --> train1.On crossing*: This ensures that whenever train 1 approaches the crossing, it will eventually pass it.

In our example the liveness condition is not satisfied: Assume that the train 2 approaches when train 1 is just leaving. The controller does not react to *approach* in state *Leaving*, hence, the transition to *Outside* appears without displaying *green* signal for train 2. The train will stop and can never reach the crossing.

The corrected finite state time machine model of the controller is shown in Figure 5.

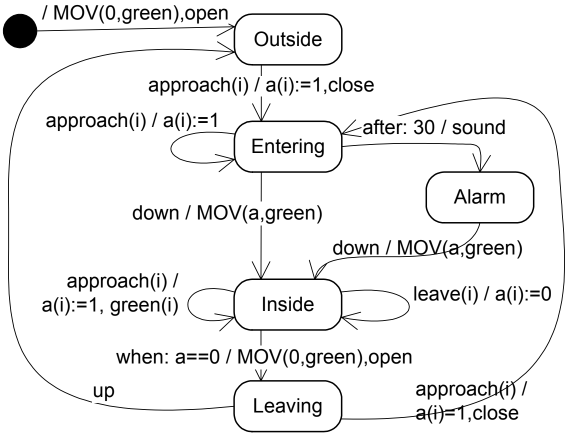


Figure 5: The corrected model of the controller.

5 CODE GENERATION

The semantics of a PLC program is defined within the reference model by the semantics of its programming language (IEC, 1993), e.g. ladder diagram or structured text. The behavior of a finite state time machine has been defined in Section 2. By that means a method for translating a high level abstract model of finite state time machine ($S, \Sigma, \Gamma, \tau, \delta, s_0, \Omega, \omega$) into a PLC program can formally be defined in the following steps:

1. Mapping of sets Σ, Ω into the input and output signals of PLC. This can be an arbitrary one-to-one mapping (coding of symbols).
2. Mapping of the set of locations which define part of state *S* into the values of flip-flops. This can be an arbitrary one-to-one mapping (coding of states). Mapping of the variables which define the other part of the state into the variables within the memory of the PLC.
3. Mapping of set Γ into the set of timers. A separate timer with the expiration time equal to $\tau_N(t)$ is allocated for each timer symbol $t \in \Gamma$.
4. Defining the function `active_timers` consistently with function τ. This function defines the input signals of all timer blocks. The input signal of a timer block allocated for a timer $t \in \Gamma$,

is a Boolean function over the set of flip-flops used for coding of states, such that it is true in state *s* if and only if $s \in \tau_S(t)$.

5. Defining function `next_state` consistently with function δ. This function defines the set and reset signals of flip-flops, which have been used for coding of states. The signal to set (reset) a flip-flop is a Boolean function over the set of flip-flops, input signals of PLC and output signal of timer blocks, such that it is true if and only if this flip-flop is set (reset) in the next state of FSTM.
6. Defining function `count_output` consistently with function ω. This function defines the values of output signals of PLC. The value of an output signal is a Boolean function over the set of flip-flops, such that it is true if and only if this output signal is set in the current state of FSTM.

Example. To capture four trains within the crossing, we need four *approach* and four *leave* input signals from trains, plus two *up* and *down* input signals from the gate (Figure 5). There are four *green* signals output to semaphores, two signals *open* and *close* to the gate and a *sound* output signal. Any combination of the input (output) signals corresponds to an input (output) symbol. PLC controller stores the locations as states of its internal flip-flops. At least three flip-flops are needed. A selected coding for states and output signals of the controller is shown in Table 1.

Table 1: The coding of states and output signals.

M1	*M2*	*M3*	*a[i]*	State	*close*	*open*	*green(i)*	*sound*
0	0	0	0	*Outside*	0	0	0	0
0	1	0	a(i)	*Entering*	1	0	0	0
1	1	0	a(i)	*Inside*	0	0	a(i)	0
1	0	0	0	*Leaving*	0	1	0	0
0	1	1	a(i)	*Alarm*	1	0	0	1

The program for PLC is a ladder diagram (IEC, 1993) consisting of a sequence of lines, each of which describes a Boolean expression to set or reset a flip-flop or an output signal, to activate a timer, or to call a function block to operate a variable, according to the values of input signals, states of flip-flops, variables and timers. The expressions reflect the coding of locations and implement the functions `active_timers`, `next_state` and `count_output` described in Section 2. An example is shown in Figure 6, which presents the transitions from *Entering* to *Alarm* and from *Entering* to *Inside* (Figure 5). M11 and M13 are auxiliary flip-flops, which mirror the main flip-flops M1 and M3, in order to assure atomicity of the transitions.

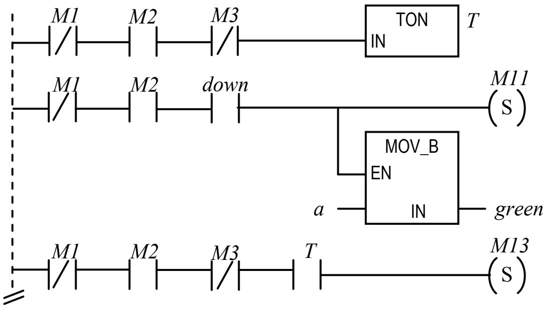


Figure 6: A fragment of the ladder diagram program for the railroad crossing controller.

6 CONCLUSIONS

A method is described for the specification, verification and automatic generation of code for PLC controllers. The advantages of the method are intuitive modeling by means of a widely accepted UML state machine, and a potential for automatic verification and implementation of the model.

A tool which implements the steps of the method has been implemented and verified on small scale examples. The verification included experiments in a lab equipped with a few process models and a set of S7 PLC controllers from Siemens.

REFERENCES

Alur R., Dill D., 1996. Automata-theoretic verification of real-time systems. In Formal Methods for Real-Time Computing, Trends in Software Series, John Wiley.

Behrmann G., David A., Larsen K.G, 2004. A Tutorial on Uppaal, Aalborg University.

Dierks, H., 1997. PLC-Automata: A New Class of Implementable Real-Time Automata. LNCS 1231. Springer, Berlin.

IEC, 1993. Programmable controllers – part 3: Programming languages.

Kaynar D.K., Lynch N.A., Segala R., Vaandrager F.W., 2006. The Theory of Timed I/O Automata. Synthesis Lecture on Computer Science, Morgan & Claypool.

Krcal P., Mokrushin L., Thiagarajan P.S., Wang Yi. 2004. Timed vs. Time Triggered Automata. LNCS 3170, Springer-Verlag, Heidelberg.

OMG, 2005. Unified Modelling Language: Superstructure, version 2.0.

Sacha K., 2007. Translatable Finite State Time Machine. LNCS 4745, Springer, Berlin.

Sacha K., 2008. Model-Based Implementation of Real-Time Systems. LNCS 5219, Springer, Berlin.

COMPUTATIONAL ASPECTS FOR RECURSIVE FRISCH SCHEME SYSTEM IDENTIFICATION

Jens G. Linden, Tomasz Larkowski and Keith J. Burnham
Control Theory and Applications Centre, Coventry University, Priory Street, CV1-5FB Coventry, U.K.
j.linden@coventry.ac.uk

Keywords: System identification, Errors-in-variables, Recursive identification.

Abstract: The implementation of a recursive algorithm for the estimation of parameters of a linear single-input single-output errors-in-variables system is re-considered. The objective is to reduce the computational complexity in order to reduce the computation time per recursion, which, in turn, will allow a wider applicability of the recursive algorithm. The technique of stationary iterative methods for least squares is utilised, in order to reduce the complexity from cubic to quadratic order with respect to the model parameters to be estimated. A numerical simulation underpins the theoretically obtained results.

1 INTRODUCTION

In the case of linear time-invariant (LTI) errors-in-variables (EIV) models not only the output signals of the system, but also the input signals are assumed to be corrupted by additive measurement noise (Söderström, 2007b). An EIV model representation can be advantageous, if the aim is to gain a better understanding of the underlying process rather than prediction. One interesting approach for the identification of dynamical systems within this framework is the so-called Frisch scheme (Beghelli et al., 1990; Söderström, 2007a), which yields estimates of the model parameters as well as the measurement noise variances. Recently, recursive Frisch scheme algorithms have been developed in a series of papers by the authors (Linden et al., 2008b; Linden et al., 2007; Linden et al., 2008a). This paper considers a fast implementation of the algorithm presented in (Linden et al., 2008a), which reduces the computational complexity from cubic to quadratic order with respect to the model parameters to be estimated, hence allowing a wider range of applicability of the proposed algorithm.

The paper is outlined as follows. The problem of EIV system identification is formulated in Section 2, where the required notation is also introduced. The Frisch scheme, being one particular EIV system identification approach is reviewed in Section 3, where non-recursive and recursive implementations are discussed. Section 4 develops the novel algorithm which reduces the computational complexity from cubic to quadratic order, whilst Section 5 presents numerical examples. Section 6 contains concluding remarks as well as direction for further work.

2 PROBLEM STATEMENT

A discrete-time, LTI single-input single-output (SISO) EIV system is considered, which is defined by

$$A(q^{-1})y_{0_i} = B(q^{-1})u_{0_i}, \tag{1}$$

where i is an integer valued time index and

$$\begin{aligned} A(q^{-1}) &\triangleq 1 + a_1 q^{-1} + \ldots + a_{n_a} q^{-n_a}, &\text{(2a)}\\ B(q^{-1}) &\triangleq b_1 q^{-1} + \ldots + b_{n_b} q^{-n_b} &\text{(2b)}\end{aligned}$$

are polynomials in the backward shift operator q^{-1}, which is defined such that $x_i q^{-1} = x_{i-1}$. The noise-free input u_{0_i} and output y_{0_i} are unknown and only the measurements

$$\begin{aligned} u_i &= u_{0_i} + \tilde{u}_i, &\text{(3a)}\\ y_i &= y_{0_i} + \tilde{y}_i &\text{(3b)}\end{aligned}$$

are available, where $\tilde{u}_i$ and $\tilde{y}_i$ denote the input and output measurement noise, respectively. Such an EIV setup is depicted in Figure 1.

The following assumptions are introduced:

A1 The dynamic system (1) is asymptotically stable, i.e. $A(q^{-1})$ has all zeros inside the unit circle.

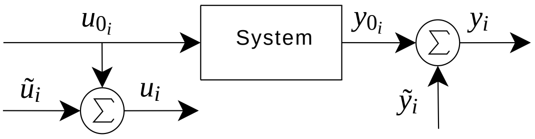


Figure 1: Errors-in-variables setup.

- **A2** All system modes are observable and controllable, i.e. $A(q^{-1})$ and $B(q^{-1})$ have no common factors.
- **A3** The polynomial degrees n_a and n_b are known *a priori* with $n_b \leq n_a$.
- **A4** The true input u_{0_i} is a zero-mean ergodic process and is persistently exciting of sufficiently high order.
- **A5** The sequences $\tilde{u}_i$ and $\tilde{y}_i$ are zero-mean, ergodic, white noises with unknown variances $\sigma_{\tilde{u}}$ and $\sigma_{\tilde{y}}$, respectively.
- **A6** The sequences $\tilde{u}_i$ and $\tilde{y}_i$ are mutually uncorrelated and uncorrelated with u_{0_i} and y_{0_i}, respectively.

A notational convention within this paper is that covariance matrices of two column vectors v_k and w_k are denoted

$$\Sigma_{vw} \triangleq E\left[v_i w_i^T\right], \qquad \Sigma_v \triangleq E\left[v_i v_i^T\right], \tag{4}$$

where $E[\cdot]$ denotes the expected value operator. In addition, vectors consisting of covariance elements are denoted

$$\xi_{vc} \triangleq E\left[v_i c_i\right] \tag{5}$$

with c_i being a scalar. The parameters are elements within a vector, which is defined by

$$\theta \triangleq \begin{bmatrix} a^T & b^T \end{bmatrix}^T = \begin{bmatrix} a_1 & \ldots & a_{n_a} & b_1 & \ldots & b_{n_b} \end{bmatrix}^T, \tag{6a}$$

$$\bar{\theta} \triangleq \begin{bmatrix} \bar{a}^T & b^T \end{bmatrix}^T = \begin{bmatrix} 1 & \theta^T \end{bmatrix}^T. \tag{6b}$$

This allows an alternative formulation of (1)-(3) given by

$$\bar{\varphi}_{0_i}^T \bar{\theta} = 0, \tag{7a}$$

$$\bar{\varphi}_i = \bar{\varphi}_{0_i} + \bar{\tilde{\varphi}}_i, \tag{7b}$$

where the regression vector is defined by

$$\begin{aligned} \varphi_i &\triangleq \begin{bmatrix} \varphi_{y_i}^T & \varphi_{u_i}^T \end{bmatrix}^T \qquad (8)\\ &\triangleq \begin{bmatrix} -y_{i-1} & \ldots & -y_{i-n_a} & u_{i-1} & \ldots & u_{i-n_b} \end{bmatrix}^T, \end{aligned}$$

$$\bar{\varphi}_i \triangleq \begin{bmatrix} \bar{\varphi}_{y_i}^T & \varphi_{u_i}^T \end{bmatrix}^T \triangleq \begin{bmatrix} -y_i & \varphi_i^T \end{bmatrix}^T. \tag{9}$$

The noise-free regression vectors φ_{0_i}, $\bar{\varphi}_{0_i}$ and the vectors containing the noise contributions $\tilde{\varphi}_i$, $\bar{\tilde{\varphi}}_i$ are defined in a similar manner. The EIV identification problem is now stated as:

Problem 1. *Given an increasing number of k samples of noisy input-output data $\{u_1, y_1, \ldots, u_k, y_k\}$, determine an estimate of the augmented parameter vector*

$$\begin{aligned} \vartheta &\triangleq \begin{bmatrix} \theta^T & \sigma^T \end{bmatrix}^T \\ &= \begin{bmatrix} a_1 & \ldots & a_{n_a} & b_1 & \ldots & b_{n_b} & \sigma_{\tilde{y}} & \sigma_{\tilde{u}} \end{bmatrix}^T. \qquad (10) \end{aligned}$$

Throughout this paper, the convention is made that estimated quantities are marked by a ˆ whilst time dependent quantities have a sub- or superscript k, e.g. $\hat{\Sigma}_{\varphi}^k$ for a sample covariance matrix corresponding to Σ_{φ}.

3 REVIEW OF THE FRISCH SCHEME

One possibility to address Problem 1 is the so called Frisch scheme (Beghelli et al., 1990; Söderström, 2006), which defines a set of admissible solutions for the estimates of the input and output measurement noise variances as well as the parameter vector. In order to single out one particular solution, different model selection criteria have been proposed within the literature, leading to different variants of the Frisch scheme (Hong et al., 2007). The criterion which is considered here is the Yule-Walker (YW) model selection criterion described in (Diversi et al., 2006) and the corresponding Frisch scheme algorithm is denoted Frisch-YW.

3.1 Non-recursive Frisch Scheme

The estimates of the (non-recursive) Frisch-YW are characterised by the input measurement noise variance $\sigma_{\tilde{u}}$, whose estimate, denoted $\hat{\sigma}_{\tilde{u}}^k$, is obtained by the nonlinear set of equations (Beghelli et al., 1990; Diversi et al., 2006)

$$\theta = \left(\hat{\Sigma}_{\varphi}^k - \Sigma_{\tilde{\varphi}}(\sigma)\right)^{-1} \hat{\xi}_{\varphi y}^k, \tag{11a}$$

$$\sigma_{\tilde{y}} = \lambda_{\min}\left(\hat{A}_k\right), \tag{11b}$$

$$\hat{\sigma}_{\tilde{u}}^k = \arg\min_{\sigma_{\tilde{u}}} V_k, \tag{11c}$$

where

$$\Sigma_{\tilde{\varphi}}(\sigma) = \begin{bmatrix} \sigma_{\tilde{y}} I_{n_a} & 0 \\ 0 & \sigma_{\tilde{u}} I_{n_b} \end{bmatrix}, \tag{12a}$$

$$\hat{A}_k \triangleq \hat{\Sigma}_{\bar{\varphi}_y}^k - \hat{\Sigma}_{\bar{\varphi}_y \varphi_u}^k \left[\hat{\Sigma}_{\varphi_u}^k - \sigma_{\tilde{u}} I_{n_b}\right]^{-1} \hat{\Sigma}_{\varphi_u \bar{\varphi}_y}^k, \tag{12b}$$

$$V_k = \frac{1}{2}\|r_k(\theta)\|_2^2 = \frac{1}{2}\|\hat{\Sigma}_{\zeta\varphi}^k \theta - \hat{\xi}_{\zeta y}^k\|_2^2, \tag{12c}$$

and $\lambda_{\min}$ denotes the minimum eigenvalue operator. The instrument vector ,denoted ζ_k, is defined by

$$\zeta_k = \begin{bmatrix} u_{k-n_b-1} & \cdots & u_{k-nb-n_\zeta} \end{bmatrix}^T \tag{13}$$

where $n_\zeta \geq n_a + n_b + 1$ denotes the number of instruments which is user specified. The quantity $r_k(\theta)$ denotes the nonlinear least squares residual corresponding to a certain θ. Once the $\sigma_{\tilde{u}}$ has been estimated, this value is substituted in (11b) and (11a), in order to obtain $\hat{\sigma}_{\tilde{y}}^k$ and $\hat{\theta}_k$, respectively. Note that (11a)-(11b) form the core of the Frisch scheme, whilst (11c) is the YW model selection criterion with the corresponding YW cost function, denoted V_k. Also note that $\hat{\theta}_k$ depends on $\hat{\sigma}_{\tilde{u}}^k$ and $\hat{\sigma}_{\tilde{y}}^k$, where the latter is also a function of $\hat{\sigma}_{\tilde{u}}^k$ defined by (11b), hence, V_k is nonlinear in $\hat{\sigma}_{\tilde{u}}^k$.

3.2 Recursive Frisch Scheme

Update of θ. The recursive Frisch scheme presented in (Linden et al., 2008b) is based on the iterative/recursive bias compensating least squares (RBCLS) approach (Sagara and Wada, 1977; Söderström, 2007b; Zheng and Feng, 1989). Assuming the noise covariances have already been obtained, the parameter vector is computed via

$$\hat{\theta}_k = \hat{\theta}_k^{\text{LS}} + P_k \Sigma_{\tilde{\varphi}}(\hat{\sigma}_k)\hat{\theta}_{k-1}, \tag{14}$$

where $\hat{\theta}_k^{\text{LS}}$ and P_k are the least squares (LS) estimate and corresponding (scaled) error covariance matrix, respectively. Both quantities are computed via the well known recursive least squares (RLS) algorithm (Ljung, 1999)

$$\hat{\theta}_k^{\text{LS}} = \hat{\theta}_{k-1}^{\text{LS}} + L_k\left(y_k - \varphi_k^T\hat{\theta}_{k-1}^{\text{LS}}\right), \tag{15a}$$

$$L_k = \frac{P_{k-1}\varphi_k}{\varphi_k^T P_{k-1}\varphi_k + \frac{1-\gamma_k}{\gamma_k}}, \tag{15b}$$

$$P_k = \frac{1}{1-\gamma_k}\left(P_{k-1} - \frac{P_{k-1}\varphi_k\varphi_k^T P_{k-1}}{\varphi_k^T P_{k-1}\varphi_k + \frac{1-\gamma_k}{\gamma_k}}\right). \tag{15c}$$

The quantity P_k is scaled such that $P_k = [\hat{\Sigma}_\varphi^k]^{-1}$, whilst the scaling factor γ_k is chosen to be $1-\lambda$ in the case of exponential forgetting, with λ being the forgetting factor.

Update of $\sigma_{\tilde{y}}$. For the determination of $\sigma_{\tilde{y}}$, a conjugate gradient subspace tracking algorithm (cf. (Feng and Owen, 1996)) has been utilised in (Linden et al., 2008b). In order to reduce the computational complexity from cubic to quadratic complexity[1], an approximate algorithm based on the Rayleigh quotient has been proposed in (Linden et al., 2007). This leads to

$$\hat{\theta}_{k-\frac{1}{2}} = \hat{\theta}_k^{\text{LS}} + P_k \begin{bmatrix} \hat{\sigma}_{\tilde{y}}^{k-1} I_{n_a} & 0 \\ 0 & \hat{\sigma}_{\tilde{u}}^k I_{n_b} \end{bmatrix} \hat{\theta}_{k-1}, \tag{16a}$$

$$\hat{\sigma}_{\tilde{y}}^k = \frac{\hat{\bar{a}}_{k-\frac{1}{2}}^T}{\hat{\bar{a}}_{k-\frac{1}{2}}^T \hat{\bar{a}}_{k-\frac{1}{2}}}\left(\hat{\Sigma}_{\varphi_y}^k \hat{\bar{a}}_{k-\frac{1}{2}} + \hat{\Sigma}_{\varphi_y\varphi_u}^k \hat{b}_{k-\frac{1}{2}}\right), \tag{16b}$$

where $\hat{\theta}_{k-\frac{1}{2}}$ denotes an intermediate parameter estimate, which makes use of the most recent estimate of $\sigma_{\tilde{u}}$ (which is determined before the update of $\hat{\sigma}_{\tilde{y}}^k$ takes place).

Update of $\sigma_{\tilde{u}}$. For the update of the input measurement noise variance $\sigma_{\tilde{u}}$, a steepest-gradient algorithm has been proposed in (Linden et al., 2008b). Recently, an alternative approach for the (approximate) minimisation of (12c) has been suggested in (Linden et al., 2008a). There, the cost function V_k is modified by replacing θ in (12c), which is nonlinear in $\sigma_{\tilde{u}}$ due to (11a) and (11b), by the approximation $L_\theta(\hat{\vartheta}_{k-1})$, which is obtained by making use of linearisations of (11a) and (11b) around the latest estimates $\hat{\vartheta}_{k-1}$. These linearisations have been developed in (Söderström, 2007a) and are given by

$$\begin{aligned}\hat{\theta}_k \approx L_\theta(\hat{\vartheta}_{k-1}) \triangleq \hat{\theta}_{k-1} + \left(\hat{\Sigma}_\varphi^k - \Sigma_{\tilde{\varphi}}(\hat{\sigma}_{k-1})\right)^{-1} \\ \times\left(\hat{\xi}_{\varphi y}^k - \hat{\Sigma}_\varphi^k\hat{\theta}_{k-1} + \begin{bmatrix}\hat{\sigma}_{\tilde{y}}^k a_{k-1} \\ \hat{\sigma}_{\tilde{u}}^k b_{k-1}\end{bmatrix}\right),\end{aligned} \tag{17a}$$

$$\hat{\sigma}_{\tilde{y}}^k \approx L_{\sigma_{\tilde{y}}}(\hat{\vartheta}_{k-1}) \triangleq \hat{\sigma}_{\tilde{y}}^{k-1} + \frac{\hat{b}_{k-1}^T\hat{b}_{k-1}}{\hat{\bar{a}}_{k-1}^T\hat{\bar{a}}_{k-1}}\left(\hat{\sigma}_{\tilde{u}}^{k-1} - \hat{\sigma}_{\tilde{u}}^k\right). \tag{17b}$$

For a convenient notation, introduce

$$\begin{aligned}\iota(\hat{\vartheta}_{k-1}) \triangleq \hat{\xi}_{\varphi y}^k - \hat{\Sigma}_\varphi^k\hat{\theta}_{k-1} \\ + \left[\hat{\sigma}_{\tilde{y}}^{k-1} + \frac{\hat{b}_{k-1}^T\hat{b}_{k-1}}{\hat{\bar{a}}_{k-1}^T\hat{\bar{a}}_{k-1}}\hat{\sigma}_{\tilde{u}}^{k-1}\right]\begin{bmatrix}\hat{a}_{k-1} \\ 0\end{bmatrix},\end{aligned} \tag{18a}$$

$$\kappa(\hat{\vartheta}_{k-1}) \triangleq \begin{bmatrix} -\frac{\hat{b}_{k-1}^T\hat{b}_{k-1}}{\hat{\bar{a}}_{k-1}^T\hat{\bar{a}}_{k-1}}\hat{a}_{k-1} \\ \hat{b}_{k-1}\end{bmatrix}, \tag{18b}$$

$$\Sigma_{\varphi_0}(\hat{\sigma}_{k-1}) \triangleq \hat{\Sigma}_\varphi^k - \Sigma_{\tilde{\varphi}}(\hat{\sigma}_{k-1}). \tag{18c}$$

Using this notation, the quantity $\sigma_{\tilde{y}}$ given by (17b) can be eliminated in (17a) yielding a linear expression for θ which only depends on $\hat{\sigma}_{\tilde{u}}^k$

$$\begin{aligned}L_\theta(\hat{\vartheta}_{k-1}) = \hat{\theta}_{k-1} + \Sigma_{\varphi_0}^{-1}(\hat{\sigma}_{k-1})\iota(\hat{\vartheta}_{k-1}) \\ + \Sigma_{\varphi_0}^{-1}(\hat{\sigma}_{k-1})\kappa(\hat{\vartheta}_{k-1})\hat{\sigma}_{\tilde{u}}^k.\end{aligned} \tag{19}$$

[1] The complexity with respect to the number of parameters to be estimated is considered. Note that the conjugate gradient algorithm in (Linden et al., 2008b) is of cubic order due to the inverse computation within the Schur complement (12b) and not due to the subspace tracking algorithm.

Substituting θ in (12c) with $L_\theta(\hat{\vartheta}_{k-1})$ allows the approximate cost function to be defined

$$V_k^{\text{lin}} \triangleq \frac{1}{2}\left\|r_k\left(L_\theta(\hat{\vartheta}_{k-1})\right)\right\|_2^2, \tag{20}$$

which can be minimised analytically at each time instance k. Differentiating with respect to θ and setting equal to zero gives

$$\hat{\sigma}^k_{\tilde{u}} = \frac{J^T(\hat{\sigma}^k_{\tilde{u}})\left(\hat{\Sigma}^k_{\zeta\varphi}\left[\hat{\theta}_{k-1} + \Sigma^{-1}_{\varphi_0}(\hat{\sigma}_{k-1})\iota(\hat{\vartheta}_{k-1})\right] - \hat{\xi}^k_{\zeta y}\right)}{-J^T(\hat{\sigma}^k_{\tilde{u}})\hat{\Sigma}^k_{\zeta\varphi}\Sigma^{-1}_{\varphi_0}(\hat{\sigma}_{k-1})\kappa(\hat{\vartheta}_{k-1})}, \tag{21}$$

where the Jacobian $J(\hat{\sigma}^k_{\tilde{u}})$ is given by

$$J(\hat{\sigma}^k_{\tilde{u}}) = \hat{\Sigma}^k_{\zeta\varphi}\frac{dL_\theta}{d\hat{\sigma}^k_{\tilde{u}}}, \tag{22}$$

whilst the total derivative of L_θ is obtained from (19) as

$$\frac{dL_\theta}{d\hat{\sigma}^k_{\tilde{u}}} = \Sigma^{-1}_{\varphi_0}(\hat{\sigma}_{k-1})\kappa(\hat{\vartheta}_{k-1}). \tag{23}$$

The resulting algorithm, which consists of (14)-(16) and (21) is referred to as recursive Frisch scheme (RFS) within the subsequent development.

4 FAST RECURSIVE FRISCH SCHEME ALGORITHM

It is observed that for the computation of the input measurement noise variance in (21), the matrix Σ_{φ_0} is required to be inverted. A matrix inversion is generally of cubic complexity with respect to its dimension, which is here equal to $n_a + n_b$, the number of parameters to be estimated. Indeed, this matrix inversion is the bottleneck within the RFS approach described in Section 3.2, since the remaining operations are only of quadratic complexity with respect to the model parameters. Since the intended use for such a recursive scheme lies in an online computation of the system parameters, it would certainly be attractive to reduce the computational burden of the input measurement noise computation to quadratic order. This would allow a wider application of the algorithm for cases where less computational power is available. The development of such an algorithm is the topic of this section.

4.1 First Bottleneck

The first bottleneck is due the computation of the inverse within the total derivative of L_θ, which has been given in Section 3.2 by (23). However, by making use of stationary iterative methods for solving LS problems (Björck, 1996, Chapter 7), Equation (23) can be re-expressed as

$$\hat{\Sigma}^k_\varphi\frac{dL_\theta}{d\hat{\sigma}^k_{\tilde{u}}} - \Sigma_{\tilde{\varphi}}(\hat{\sigma}_{k-1})\frac{dL_\theta}{d\hat{\sigma}^k_{\tilde{u}}} = \kappa(\hat{\vartheta}_{k-1}), \tag{24}$$

where the matrix splitting is given naturally by (18c). An iterative/recursive way to compute $dL_\theta/d\hat{\sigma}^k_{\tilde{u}}$ could therefore be given by

$$L_\theta^{k'} \triangleq P_k\left[\kappa(\hat{\vartheta}_{k-1}) + \Sigma_{\tilde{\varphi}}(\hat{\sigma}_{k-1})L_\theta^{k-1'}\right], \tag{25}$$

where $L_\theta^{k'}$ denotes the recursively computed derivative and $P_k = [\hat{\Sigma}^k_\varphi]^{-1}$ is given by the matrix inversion lemma of the RLS algorithm, which is already computed for the determination of $\hat{\theta}_k$.

4.2 Second Bottleneck

The second bottleneck is due to the matrix inverse within the computation of (19), therefore, an (approximate) recursive expression for $L_\theta(\hat{\vartheta}_{k-1})$ is required, which is of quadratic complexity only. Firstly, introduce the notation $L_\theta(\hat{\vartheta}_{k-1}) \triangleq L^k_\theta$, where the index k is chosen to reflect the fact that L^k_θ corresponds to the linearisation at time instance k (although it depends on the estimate $\hat{\vartheta}_{k-1}$ with time index $k-1$). Secondly, assume that all past $\hat{\theta}_k$ have been computed using the expression (19), which means that $\hat{\theta}_{k-1}$ can be replaced with L^{k-1}_θ in (19). Thirdly, from (18a) and (18b) it holds

$$\begin{aligned}\iota(\hat{\vartheta}_{k-1}) + \kappa(\hat{\vartheta}_{k-1})\hat{\sigma}^k_{\tilde{u}} = \hat{\xi}^k_{\varphi y} - \hat{\Sigma}^k_\varphi\hat{\theta}_{k-1} + \begin{bmatrix}\hat{a}_{k-1}\\0\end{bmatrix}\hat{\sigma}^{k-1}_{\tilde{y}}\\ + \begin{bmatrix}\frac{\hat{b}^T_{k-1}\hat{b}_{k-1}}{\hat{a}^T_{k-1}\hat{a}_{k-1}}\hat{a}_{k-1}\\0\end{bmatrix}\hat{\sigma}^{k-1}_{\tilde{u}} + \begin{bmatrix}-\frac{\hat{b}^T_{k-1}\hat{b}_{k-1}}{\hat{a}^T_{k-1}\hat{a}_{k-1}}\hat{a}_{k-1}\\0\end{bmatrix}\hat{\sigma}^k_{\tilde{u}}\\ + \begin{bmatrix}0\\\hat{b}_{k-1}\end{bmatrix}\hat{\sigma}^k_{\tilde{u}},\end{aligned} \tag{26}$$

and by assuming that $\hat{\sigma}^k_{\tilde{u}} \approx \hat{\sigma}^{k-1}_{\tilde{u}}$, $\hat{\sigma}^k_{\tilde{y}} \approx \hat{\sigma}^{k-1}_{\tilde{y}}$ and using $\hat{\theta}_{k-1} = L^{k-1}_\theta$, one obtains

$$\begin{aligned}\iota(\hat{\vartheta}_{k-1}) + \kappa(\hat{\vartheta}_{k-1})\hat{\sigma}^k_{\tilde{u}} \approx \hat{\xi}^k_{\varphi y} - \hat{\Sigma}^k_\varphi L^{k-1}_\theta\\ + \begin{bmatrix}\hat{\sigma}^k_{\tilde{y}}I_{n_a} & 0\\0 & \hat{\sigma}^k_{\tilde{u}}I_{n_b}\end{bmatrix}L^{k-1}_\theta.\end{aligned} \tag{27}$$

Finally, by substituting (18c) and (27) into (19), it holds

$$\begin{aligned}\left[\hat{\Sigma}^k_\varphi - \Sigma_{\tilde{\varphi}}(\hat{\sigma}_{k-1})\right]L^k_\theta = \left[\hat{\Sigma}^k_\varphi - \Sigma_{\tilde{\varphi}}(\hat{\sigma}_{k-1})\right]L^{k-1}_\theta\\ + \hat{\xi}^k_{\varphi y} - \hat{\Sigma}^k_\varphi L^{k-1}_\theta + \Sigma_{\tilde{\varphi}}(\hat{\sigma}_k)L^{k-1}_\theta,\end{aligned} \tag{28}$$

which simplifies to

$$\left[\hat{\Sigma}_{\varphi}^{k} - \Sigma_{\tilde{\varphi}}(\hat{\sigma}_{k-1})\right] L_{\theta}^{k} = -\Sigma_{\tilde{\varphi}}(\hat{\sigma}_{k-1}) L_{\theta}^{k-1} + \hat{\xi}_{\varphi y}^{k} + \Sigma_{\tilde{\varphi}}(\hat{\sigma}_{k}) L_{\theta}^{k-1}. \tag{29}$$

Thus, by using $L_{\theta}^{k-1} \approx L_{\theta}^{k}$, a recursive computation of the linearised θ-equation (17a) is given by

$$L_{\theta}^{k} \approx [\hat{\Sigma}_{\varphi}^{k}]^{-1} \hat{\xi}_{\varphi y}^{k} + [\hat{\Sigma}_{\varphi}^{k}]^{-1} \Sigma_{\tilde{\varphi}}(\hat{\sigma}_{k}) L_{\theta}^{k-1}, \tag{30}$$

which interestingly is, indeed, the RBCLS algorithm given in (14) (i.e. simply replace $\hat{\theta}_k$ with L_{θ}^{k} in (14)). Since the recursive computation of L_{θ}^{k} is identical to the RBCLS computation of $\hat{\theta}_k$, the latter, more familiar, notation can be utilised. Substituting the linearised $\lambda_{\min}$-equation (17b), the RBCLS equation becomes

$$\hat{\theta}_k = \hat{\theta}_k^{\mathrm{LS}} + P_k \begin{bmatrix} \hat{\sigma}_{\tilde{y}}^{k} \hat{a}_{k-1} \\ \hat{\sigma}_{\tilde{u}}^{k} b_{k-1} \end{bmatrix}$$

$$\Leftrightarrow \quad \hat{\theta}_k = \hat{\theta}_k^{\mathrm{LS}} + P_k \begin{bmatrix} 0 \\ \hat{b}_{k-1} \end{bmatrix} \hat{\sigma}_{\tilde{u}}^{k} + P_k \begin{bmatrix} \hat{a}_{k-1} \\ 0 \end{bmatrix} \times \left[\hat{\sigma}_{\tilde{y}}^{k-1} + \frac{\hat{b}_{k-1}^{T} \hat{b}_{k-1}}{\hat{a}_{k-1}^{T} \hat{a}_{k-1}} \hat{\sigma}_{\tilde{u}}^{k-1} - \frac{\hat{b}_{k-1}^{T} \hat{b}_{k-1}}{\hat{a}_{k-1}^{T} \hat{a}_{k-1}} \hat{\sigma}_{\tilde{u}}^{k} \right], \tag{31}$$

which simplifies to

$$\hat{\theta}_k = P_k \bar{\iota}(\hat{\vartheta}_{k-1}) + P_k \kappa(\hat{\vartheta}_{k-1}) \hat{\sigma}_{\tilde{u}}^{k}, \tag{32}$$

where $\kappa(\hat{\vartheta}_{k-1})$ is defined by (18b) and

$$\bar{\iota}(\hat{\vartheta}_{k-1}) \triangleq \hat{\xi}_{\varphi y}^{k} + \begin{bmatrix} \hat{a}_{k-1} \\ 0 \end{bmatrix} \left[\hat{\sigma}_{\tilde{y}}^{k-1} + \frac{\hat{b}_{k-1}^{T} \hat{b}_{k-1}}{\hat{a}_{k-1}^{T} \hat{a}_{k-1}} \hat{\sigma}_{\tilde{u}}^{k-1} \right]. \tag{33}$$

4.3 Fast Update of $\hat{\sigma}_{\tilde{u}}^{k}$

Using the previous results, a fast implementation for the update of $\hat{\sigma}_{\tilde{u}}^{k}$ can be realised. With the Jacobian at time instance k being given by (cf. (22))

$$J_k \triangleq \hat{\Sigma}_{\zeta\varphi}^{k} L_{\theta}^{k'}, \tag{34}$$

it is therefore possible to compute $\hat{\sigma}_{\tilde{u}}^{k}$ as

$$0 = J_k^T \left[\hat{\Sigma}_{\zeta\varphi}^{k} \hat{\theta}_k - \hat{\xi}_{\zeta y}^{k} \right] \tag{35}$$

and by substituting $\hat{\theta}_k$ given in (32), the fast update for $\hat{\sigma}_{\tilde{u}}^{k}$ is finally given by

$$\hat{\sigma}_{\tilde{u}}^{k} = \frac{J_k^T \left[\hat{\xi}_{\zeta y}^{k} - \hat{\Sigma}_{\zeta\varphi}^{k} P_k \bar{\iota}(\hat{\vartheta}_{k-1}) \right]}{J_k^T \hat{\Sigma}_{\zeta\varphi}^{k} P_k \kappa(\hat{\vartheta}_{k-1})}. \tag{36}$$

Note that only matrix vector multiplications are required for the fast computation of $\hat{\sigma}_{\tilde{u}}^{k}$, hence the computational effort is reduced towards quadratic complexity. The fast RFS algorithm, which consists of (14)-(16) and (36) is referred to as FRFS within the subsequent development.

5 NUMERICAL EXAMPLES

It is of interest to compare the RFS estimates with those obtained by the FRFS and also to compare the computation time of both algorithms.

5.1 Estimation of $\sigma_{\tilde{u}}$

A LTI SISO system with $n_a = n_b = 2$ and given by

$$\theta = \begin{bmatrix} -1.5 & 0.7 & 1 & 0.5 \end{bmatrix}^T \tag{37a}$$

$$\sigma = \begin{bmatrix} 2.1 & 0.1 \end{bmatrix}^T \tag{37b}$$

is simulated for 1000 samples using a zero mean, white and Gaussian distributed input signal of unity variance. The RFS and FRFS algorithms are applied to estimate ϑ using $n_\zeta = n_a + n_b + 1$, whilst $\lambda = 1$ is chosen (i.e. no forgetting). The estimates of $\sigma_{\tilde{u}}$ and $\sigma_{\tilde{y}}$ are projected into the intervals $[0, \sigma_{\tilde{u}}^{\max}]$ and $[0, \sigma_{\tilde{u}}^{\max}]$, where the maximal admissible values for the input and output measurement noise variances are chosen to be $\sigma_{\tilde{u}}^{\max} = 2\sigma_{\tilde{u}} = 0.2$ and $\sigma_{\tilde{y}}^{\max} = 2\sigma_{\tilde{y}} = 4.2$, respectively. The estimates of $\sigma_{\tilde{u}}$ are compared in Figure 2. Here it is observed that the projection facility (which sets $\hat{\sigma}_{\tilde{u}}^{k} = \hat{\sigma}_{\tilde{u}}^{k-1}$ if the estimate is not within the specified interval) seems to be more often active for the fast algorithm (see around $k = 420$). After approximately 500 recursions, however, the FRFS estimate is barely distinguishable from the RFS, although the FRFS solution seems to be slightly more erratic. Hence, at least in the example considered here, the FRFS appears to be able to approximate the estimate of $\sigma_{\tilde{u}}$ obtained by the more computationally demanding RFS algorithm.

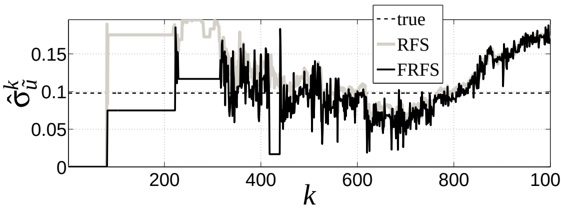


Figure 2: Estimates of $\sigma_{\tilde{u}}$ for using the RFS, and FRFS.

5.2 Comparison of Computation Time

Naturally, it is of major interest to compare the computation time per recursion of the FRFS algorithm with that of the RFS scheme. Therefore, the algorithms are applied to systems with an incrementally increasing model order $m = n_a = n_b = 1, \ldots, 30$ and the computation time per single recursion is recorded for each identification task. The results are presented in Figure 3, which clearly shows the relative reduction of computational complexity for the FRFS approach. For a model order of $m = 30$, the RFS requires around 3.5ms, whilst the FRFS requires less than 2.0ms. The fact that the slope of the curve corresponding to the FRFS algorithm is lower than that of the RFS approach illustrates that the computational complexity is reduced from cubic to quadratic order; this underpins the theoretical results obtained in this paper.

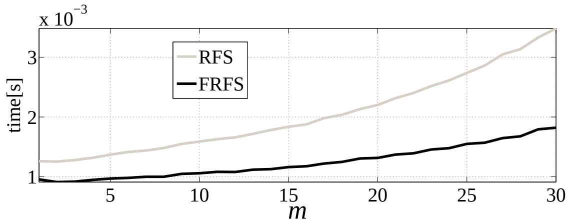


Figure 3: Computation time per single recursion with increasing model order m.

6 CONCLUSIONS AND FURTHER WORK

The Frisch scheme for the identification of linear time-invariant single-input single-output errors-in-variables systems has been reviewed. The well-known non-recursive case as well as a recently developed recursive algorithm has been discussed. Since the latter is of cubic computational complexity with respect to the number of parameters to be estimated, several approximations have been introduced, in order to reduce the complexity from cubic to quadratic order. This theoretical result is in agreement with the measured computation time which has been obtained for a numerical simulation. This simulation has also shown that the fast algorithm is able to approximate the solution of the computationally more demanding algorithm satisfactorily.

Further work could concern the convergence properties of the recursive algorithm.

REFERENCES

Beghelli, S., Guidorzi, R. P., and Soverini, U. (1990). The Frisch scheme in dynamic system identification. *Automatica*, 26(1):171–176.

Björck, Å. (1996). *Numerical Methods for Least Squares Problems*. SIAM, Philadelphia.

Diversi, R., Guidorzi, R., and Soverini, U. (2006). Yule-Walker equations in the Frisch scheme solution of errors-in-variables identification problems. In *Proc. 17th Int. Symp. on Math. Theory of Networks and Systems*, pages 391–395.

Feng, Y. T. and Owen, D. R. J. (1996). Conjugate gradient methods for solving the smallest eigenpair of large symmetric eigenvalue problems. *Int. J. for Numerical Methods in Engineering*, 39:2209–2229.

Hong, M., Söderström, T., Soverini, U., and Diversi, R. (2007). Comparison of three Frisch methods for errors-in-variables identification. Technical Report 2007-021, Uppsala University, Uppsala, Sweden.

Linden, J. G., Larkowski, T., and Burnham, K. J. (2008a). An improved recursive Frisch scheme identification algorithm. In *Proc. 19th Int. Conf. on Systems Engineering*, pages 65–70, Las Vegas, USA.

Linden, J. G., Vinsonneau, B., and Burnham, K. J. (2007). Fast algorithms for recursive Frisch scheme system identification. In *Proc. CD-ROM 22nd IAR & ACD Workshop*, Grenoble, France.

Linden, J. G., Vinsonneau, B., and Burnham, K. J. (2008b). Gradient-based approaches for recursive Frisch scheme identification. In *Preprints of the 17th IFAC World Congress*, pages 1390–1395, Seoul, Korea.

Ljung, L. (1999). *System Identification - Theory for the user*. PTR Prentice Hall Information and System Sciences Series. Prentice Hall, New Jersey, 2nd edition.

Sagara, S. and Wada, K. (1977). On-line modified least-squares parameter estimation of linear discrete dynamic systems. *Int. J. Control*, 25(3):329–343.

Söderström, T. (2006). Statistical analysis of the Frisch scheme for identifying errors-in-variables systems. Technical report 2006-002, Uppsala University, Department of Information Technology, Uppsala, Sweden.

Söderström, T. (2007a). Accuracy analysis of the Frisch scheme for identifying errors-in-variables systems. *IEEE Trans. Autom. Contr.*, 52(6):985–997.

Söderström, T. (2007b). Errors-in-variables methods in system identification. *Automatica*, 43(6):939–958.

Zheng, W. X. and Feng, C. B. (1989). Unbiased parameter estimation of linear systems in the presence of input and output noise. *Int. J. of Adaptive Control and Signal Proc.*, 3:231–251.

RECURSIVE EXTENDED COMPENSATED LEAST SQUARES BASED ALGORITHM FOR ERRORS-IN-VARIABLES IDENTIFICATION

Tomasz Larkowski, Jens G. Linden and Keith J. Burnham
Control Theory and Applications Centre, Coventry University, Priory Street, Coventry, CV1 5FB, U.K.
larkowst@coventry.ac.uk

Keywords: Bias compensation, Errors-in-variables, Parameter estimation, Recursive algorithms, System identification.

Abstract: An algorithm for the recursive identification of single-input single-output linear discrete-time time-invariant errors-in-variables system models in the case of white input and coloured output noise is presented. The approach is based on a bilinear parametrisation technique which allows the model parameters to be estimated together with the auto-correlation elements of the input/output noise sequences. In order to compensate for the bias in the recursively obtained least squares estimates, the extended bias compensated least squares method is used. An alternative for the online update of the associated pseudo-inverse of the extended observation covariance matrix is investigated, namely an approach based on the matrix pseudo-inverse lemma and an approach based on the recursive extended instrumental variables technique. A Monte-Carlo simulation study demonstrates the appropriateness and the robustness against noise of the proposed scheme.

1 INTRODUCTION

The errors-in-variables (EIV) approach forms an extension of the standard output error system setup in which it is postulated that only the output measurements are uncertain. In the EIV framework all measured signals, hence, including the system input, are assumed to be contaminated with noise, see (Söderström, 2007) for the recent survey on this subject. The EIV framework can offer advantages over the classical approach, mainly when the description of the internal laws governing a system is of prime interest, e.g. application areas in chemistry, image processing, fault detection etc., see (Söderström, 2007; Markovsky and Van Huffel, 2007) for further details.

One of the EIV techniques that has been shown to be robust and to yield relatively precise estimates is the extended compensated least squares (ECLS) method. The approach is based on the extended bias compensated least squares (EBCLS) and utilises separable nonlinear least squares to solve the resulting overall identification problem. The method was first proposed in (Ekman, 2005a), which considered the case of white input and output noise sequences and subsequently extended to handle the case of coloured output noise in (Ekman et al., 2006). Further analysis, considering a generalised framework, has been carried out in (Mahata, 2007).

Alternatively, by exploiting the property that the overall optimisation problem is bilinear in the unknowns, see (Ljung, 1999), which in this case corresponds to the model parameters and the input/output noise auto-correlation elements, the principle of bilinear parametrisation can be utilised. The resulting scheme, termed here the extended bilinear parametrisation method (EBPM) involves solving iteratively two ordinary least squares problems, see (Larkowski et al., 2008) for details. Although the quality of the parameters obtained by the EBPM is comparable to the quality of the estimates yielded by the ECLS, an important distinction is that the EBPM is significantly less computationally demanding than the ECLS technique.

The bilinear parametrisation method was first utilised to solve the EIV identification problem in a recursive manner in (Ekman, 2005b) for the case of white input and output noise. It has also been exploited in (Ikenoue et al., 2008) for the case of coloured input and output noise sequences and for the purpose of offline as well as online estimation. However, in both cases the term ‘bilinear parametrisation’ has not been explicitly stated. In (Ekman, 2005b) the constructed recursive algorithm is not computationally attractive, since its complexity at each iteration is actually greater than that of the corresponding batch algorithm applied in an offline manner at each recursion. Whereas, in (Ikenoue et al., 2008) due to a special choice of the instruments, the resulting algorithm is not causal, in general, hence its recursive implementation yields delayed estimates.

In this paper a recursive realisation of the EBPM is presented for a discrete-time linear time-invariant (LTI) single-input single-output (SISO) system model in the case of the white input and coloured output noise and it is demonstrated that the above mentioned shortcomings may be avoided. The bias of the recursively calculated least squares (LS) estimator is removed at each recursion via the extended bias compensated least squares (EBCLS) technique. The online update of the pseudo-inverse of the overdetermined observation matrix is realised by considering an alternative, namely an approach based on the pseudo-inverse lemma, see (Feng et al., 2001) and an approach based on the recursive extended instrumental variables technique, see (Friedlander, 1984). The two resulting algorithms are analysed and compared with their offline counterpart via a Monte-Carlo simulation study. It is shown that the instrumental variables approach is the more preferable due to its superior robustness and improved convergence properties, in general.

2 NOTATION AND PROBLEM STATEMENT

Consider a discrete-time LTI SISO system represented by the difference equation

$$A(q^{-1})y_{0_k} = B(q^{-1})u_{0_k}, \tag{1}$$

where the polynomials $A(q^{-1})$ and $B(q^{-1})$ are given by

$$A(q^{-1}) \triangleq 1 + a_1 q^{-1} + \ldots + a_{n_a} q^{-n_a}, \tag{2a}$$

$$B(q^{-1}) \triangleq b_1 q^{-1} + \ldots + b_{n_b} q^{-n_b} \tag{2b}$$

with q^{-1} being the backward shift operator, defined by $q^{-1}x_k \triangleq x_{k-1}$. The unknown noise-free input and noise-free output signals denoted u_{0_k} and y_{0_k}, respectively, are related to the available noisy variables, denoted u_k and y_k, such that

$$u_k = u_{0_k} + \tilde{u}_k, \qquad y_k = y_{0_k} + \tilde{y}_k, \tag{3}$$

where $\tilde{u}_k$ and $\tilde{y}_k$ denote the input and output measurement noise sequences, respectively. The following standard assumptions, see e.g. (Ekman et al., 2006), are introduced:

A1 The LTI system (1) is asymptotically stable, i.e. $A(q^{-1})$ has all zeros inside the unit circle.

A2 All system modes are observable and controllable, i.e. $A(q^{-1})$ and $B(q^{-1})$ share no common factors.

A3 The system structure, i.e. n_a and n_b, is known *a priori* and $n_a \geq n_b$.

A4 The true input u_{0_k} is a zero mean, ergodic random sequence persistently exciting and of sufficiently high order, i.e. at least of order $n_a + n_b$.

A5a The additive input noise sequence $\tilde{u}_k$ of unknown variance $\sigma_{\tilde{u}}$ is an ergodic zero mean white process.

A5b The additive output noise sequence $\tilde{y}_k$ is an ergodic zero mean process characterised by an unknown auto-covariance sequence $\{r_{\tilde{y}}(0),\ r_{\tilde{y}}(1),\ \ldots\}$.

A6 The input/output noise sequences are mutually uncorrelated and uncorrelated with signals u_{0_k} and y_{0_k}.

By postulating that the output noise sequence exhibits an arbitrary degree of correlation allows for measurement sensor uncertainties to be taken into account, as well as potential disturbances in the process.

The system parameter vector is denoted

$$\theta \triangleq \begin{bmatrix} a^T & b^T \end{bmatrix}^T \in \mathcal{R}^{n_\theta} \qquad , \tag{4a}$$

$$a \triangleq \begin{bmatrix} a_1 & \ldots & a_{n_a} \end{bmatrix}^T \in \mathcal{R}^{n_a}, \tag{4b}$$

$$b \triangleq \begin{bmatrix} b_1 & \ldots & b_{n_b} \end{bmatrix}^T \in \mathcal{R}^{n_b}, \tag{4c}$$

where $n_\theta = n_a + n_b$. The extended regressor vectors for the k-th measured data are defined as

$$\bar{\varphi}_k \triangleq \begin{bmatrix} -y_k & \varphi_k^T \end{bmatrix}^T \in \mathcal{R}^{n_\theta + 1}, \tag{5a}$$

$$\bar{\varphi}_{y_k} \triangleq \begin{bmatrix} -y_k & \varphi_{y_k}^T \end{bmatrix}^T \in \mathcal{R}^{n_a + 1}, \tag{5b}$$

where

$$\varphi_k \triangleq \begin{bmatrix} \varphi_{y_k}^T & \varphi_{u_k}^T \end{bmatrix}^T \in \mathcal{R}^{n_\theta}, \tag{5c}$$

$$\varphi_{y_k} \triangleq \begin{bmatrix} -y_{k-1} \ldots -y_{k-n_a} \end{bmatrix}^T \in \mathcal{R}^{n_a}, \tag{5d}$$

$$\varphi_{u_k} \triangleq \begin{bmatrix} u_{k-1} \ldots u_{k-n_b} \end{bmatrix}^T \in \mathcal{R}^{n_b}. \tag{5e}$$

The noise contributions in the corresponding regressor vectors are denoted by a tilde, i.e. $\tilde{[\cdot]}$, whereas the noise-free signals are denoted by a zero subscript, i.e. $[\cdot]_0$. From (3) it follows that

$$\bar{\varphi}_k = \bar{\varphi}_{0_k} + \tilde{\bar{\varphi}}_k. \tag{6}$$

The notation Σ_{gd} is used as a general notion for the covariance matrix of the vectors g_k and d_k, whereas ξ_{gf} is utilised for a covariance vector with f_k being a scalar. The corresponding estimates are denoted by a hat. In addition, $0_{g \times d}$ denotes the null matrix of arbitrary dimension $g \times d$ and a single index is used in the case of a column vector as well as in the case of a square matrix, e.g. the identity matrix I_g. The autocorrelation elements, denoted $r_{\tilde{y}}(\cdot)$ are defined as

$$r_{\tilde{y}}(\tau) \triangleq E\left[\tilde{y}_k \tilde{y}_{k-\tau}\right], \tag{7}$$

where $E[\cdot]$ is the expected value operator. Introducing

$$\rho \triangleq \begin{bmatrix} \rho_y^T & \sigma_{\tilde{u}} \end{bmatrix}^T \in \mathcal{R}^{n_a+2}, \tag{8a}$$

$$\rho_y \triangleq \begin{bmatrix} r_{\tilde{y}}(0) & \dots & r_{\tilde{y}}(n_a) \end{bmatrix}^T \in \mathcal{R}^{n_a+1}, \tag{8b}$$

the dynamic identification problem in the EIV framework considered here is formulated as:

Problem 1. *(Dynamic EIV identification problem) Given N samples of the measured signals, i.e.* $\{u_k\}_{k=1}^N$ *and* $\{y_k\}_{k=1}^N$, *determine the vector*

$$\Theta \triangleq \begin{bmatrix} \theta^T & \rho^T \end{bmatrix}^T \in \mathcal{R}^{n_\theta+n_a+2}. \tag{9}$$

3 REVIEW OF APPROACHES

This section briefly reviews the EBCLS technique and the offline EBPM algorithm.

3.1 Extended Bias Compensated Least Squares

Denoting an estimate by $\hat{[\cdot]}$, a solution of the system (1)-(3) in the LS sense is given by

$$\hat{\theta}_{\text{LS}} = \hat{\Sigma}_{x\varphi}^{\dagger} \hat{\xi}_{xy}, \tag{10}$$

where $[\cdot]^{\dagger}$ is the pseudo inverse operator defined by $A^{\dagger} \triangleq (A^T A)^{-1} A^T$, $x_k \in \mathcal{R}^{n_x}$ denotes an arbitrary instrumental vector with $n_x \geq n_\theta$. Due to the measurement noise, unless the elements of x_k are uncorrelated with $\tilde{\varphi}_k$, the solution obtained is biased. In order to achieve an unbiased estimate of θ, a bias compensation procedure is required to be carried out (Söderström, 2007). This consideration yields the EBCLS estimator defined as

$$\hat{\theta}_{\text{EBCLS}} \triangleq \left(\hat{\Sigma}_{x\varphi} - \Sigma_{\tilde{x}\tilde{\varphi}}\right)^{\dagger} \left(\hat{\xi}_{xy} - \xi_{\tilde{x}\tilde{y}}\right). \tag{11}$$

Note that $\Sigma_{\tilde{x}\tilde{\varphi}}$ and $\xi_{\tilde{x}\tilde{y}}$, in general, are functions of ρ, which, in turn, will depend on the elements contained in the instrument vector x_k.

3.2 Extended Bilinear Parametrisation Method

The bilinear parametrisation method is applicable for problems that are bilinear in the parameters, see (Ljung, 1999) for details, and it is presented here in accordance with the development proposed in (Larkowski et al., 2008).

Based on the EBCLS rule given by (11) a bilinear (in the parameters) cost function can be formulated, i.e.

$$\hat{\Theta} = \arg\min_{\Theta} V(\Theta), \tag{12}$$

where

$$V(\Theta) \triangleq \left\| \hat{\xi}_{xy} - \xi_{\tilde{x}\tilde{y}} - \left(\hat{\Sigma}_{x\varphi} - \Sigma_{\tilde{x}\tilde{\varphi}}\right)\theta \right\|_2^2. \tag{13}$$

Note that the instruments x_k must be chosen such that the resulting problem is soluble, i.e. the total number of unknowns is less than or equal to the total number of equations, see (Larkowski et al., 2008) for a detailed treatment. Alternatively, utilising the bilinearity property, (13) can be re-expressed as

$$V(\Theta) = \left\| \hat{\xi}_{xy} - \hat{\Sigma}_{x\varphi}\theta - W\rho \right\|_2^2, \tag{14}$$

where $W \triangleq S_1 - S_2(\theta) \in \mathcal{R}^{n_x \times (n_a+2)}$ such that $S_1\rho \triangleq \xi_{\tilde{x}\tilde{y}}$ and $S_2(\theta)\rho \triangleq \Sigma_{\tilde{x}\tilde{\varphi}}\theta$.

It is observed that for fixed ρ (i.e. the expressions $\Sigma_{\tilde{x}\tilde{\varphi}}$ and $\xi_{\tilde{x}\tilde{y}}$) the cost function (13) is linear in θ. Analogously, for fixed θ (i.e. the matrix W) the cost function (14) is linear in ρ. Consequently, a natural approach is to treat (13) and (14) as separate LS problems, cf. (Ljung, 1999). This leads to a two-step algorithm where the LS solutions of the sub-problems defined by (13) and (14) are obtained at each iteration. Furthermore, local convergence of such algorithm is guaranteed, see (Ljung, 1999).

4 RECURSIVE EXTENDED BILINEAR PARAMETRISATION METHOD

This section presents the proposed recursive realisation of the EBPM technique, denoted REBPM. First the problem of an online update of the parameter vector is addressed. Subsequently, two approaches for updating the pseudo-inverse of the extended observation matrix are considered. Finally, the problem of calculating the input noise variance and the autocorrelation elements of the output noise is discussed.

4.1 Recursive Update of Parameter Vector

Considering (11) and by making use of (10) it follows that

$$\theta = \hat{\theta}_{LS} + \hat{\Sigma}_{x\varphi}^{\dagger}\left(\Sigma_{\tilde{x}\tilde{\varphi}}\theta - \xi_{\tilde{x}\tilde{y}}\right). \tag{15}$$

It is remarked that the expression $\hat{\Sigma}^{\dagger}_{x\varphi}\left(\Sigma_{\tilde{x}\tilde{\varphi}}\theta - \xi_{\tilde{x}\tilde{y}}\right)$ represents the bias of the LS estimator. Since the true value of θ on the right hand side of (15) is unknown, a natural approach is to utilise the most recent estimate, i.e. the previous value. This leads to the following recursive EBCLS scheme

$$\hat{\theta}^{k}_{\text{EBCLS}} = \hat{\theta}^{k}_{\text{LS}} + \left(\hat{\Sigma}^{k}_{x\varphi}\right)^{\dagger}\left(\Sigma^{k}_{\tilde{x}\tilde{\varphi}}\hat{\theta}^{k-1}_{\text{EBCLS}} - \xi^{k}_{\tilde{x}\tilde{y}}\right). \quad (16)$$

Despite that an inevitable error is introduced by assuming $\hat{\theta}^{k}_{\text{EBCLS}} \approx \hat{\theta}^{k-1}_{\text{EBCLS}}$, the above approach also known as the stationary iterative LS principle (Björck, 1996), has been successfully employed in several recursive as well as iterative algorithms.

4.2 Recursive Update of Pseudo-inverse

Considering equation (16), it is observed that a recursive update of the pseudo-inverse of $\hat{\Sigma}^{k}_{x\varphi}$ as well as of the LS estimate, i.e. $\hat{\theta}^{k}_{\text{LS}}$, is required. This problem can be tackled by two approaches described below.

Approach based on the Matrix Pseudo-inverse Lemma - REBPM$_1$. The first, i.e. direct approach is to utilise an extension of the matrix inverse lemma, namely the matrix pseudo-inverse lemma, see (Feng et al., 2001). This allows the recursive computation of the expression $\hat{\Sigma}^{\dagger}_{x\varphi}$ as well as the corresponding $\hat{\theta}_{\text{LS}}$. The algorithm can be summarised as follows:

$$\hat{\theta}^{k}_{\text{LS}} = \hat{\theta}^{k-1}_{\text{LS}} + L_k\left(y_k - \varphi^{T}_{k}\hat{\theta}^{k-1}_{\text{LS}}\right), \quad (17a)$$

$$L_k = \frac{\left(\hat{\Sigma}^{k-1}_{x\varphi}\right)^{\dagger} x_k}{k-1+\varphi^{T}_{k}\left(\hat{\Sigma}^{k-1}_{x\varphi}\right)^{\dagger} x_k}, \quad (17b)$$

$$\left(\hat{\Sigma}^{k}_{x\varphi}\right)^{\dagger} = \frac{k}{k-1}\left[\left(\hat{\Sigma}^{k-1}_{x\varphi}\right)^{\dagger} - L_k\varphi^{T}_{k}\left(\hat{\Sigma}^{k-1}_{x\varphi}\right)^{\dagger}\right], \quad (17c)$$

$$\hat{\Sigma}^{k}_{x\varphi} = \hat{\Sigma}^{k-1}_{x\varphi} + \frac{1}{k}\left(x_k\varphi^{T}_{k} - \hat{\Sigma}^{k-1}_{x\varphi}\right), \quad (17d)$$

$$\hat{\xi}^{k}_{xy} = \hat{\xi}^{k-1}_{xy} + \frac{1}{k}\left(x_k y_k - \hat{\xi}^{k-1}_{xy}\right). \quad (17e)$$

The main shortcoming of the pseudo-inverse approach when dealing with practical applications results from its relatively high sensitivity with respect to the initialisation of the pseudo-inverse of the matrix $\hat{\Sigma}^{k}_{x\varphi}$. This issue is not trivial and can lead to a divergence of the overall algorithm. In order to appropriately initialise the expression $\left(\hat{\Sigma}^{k}_{x\varphi}\right)^{\dagger}$, it is required that the pseudo-inverse is computed offline after an arbitrary number, denoted α, of measurements is taken and before the recursive algorithm commences operation.

Remark 1. *It is noted that the uniqueness of* $\left(\hat{\Sigma}_{x\varphi}\right)^{\dagger}$ *in the case of recursive approaches is not always guaranteed when utilising equations* (17)*, see (Linden, 2008) for further details. As a consequence, the corresponding estimate of* θ^{k}_{LS} *may not represent the optimal, in terms of the minimum variance, solution to the overdetermined set of equations given by* (10)*.*

Approach based on Extended Instrumental Variables - REBPM$_2$. An alternative to employing the matrix pseudo-inverse lemma, an approach based on the recursive extended instrumental variables technique, see (Friedlander, 1984), can be utilised in order to obtain, albeit indirectly, a recursive update of $\left(\hat{\Sigma}^{k}_{x\varphi}\right)^{\dagger}$. Define

$$P_k = \left[\left(\hat{\Sigma}^{k}_{x\varphi}\right)^{T}\hat{\Sigma}^{k}_{x\varphi}\right]^{-1}. \quad (18)$$

In this approach the expression P_k is updated recursively, rather than the total pseudo-inverse $\left(\hat{\Sigma}^{k}_{x\varphi}\right)^{\dagger}$. The algorithm can be summarised as:

$$\hat{\theta}^{k}_{\text{LS}} = \hat{\theta}^{k-1}_{\text{LS}} + K_k\left(v_k - \phi^{T}_{k}\hat{\theta}^{k-1}_{\text{LS}}\right), \quad (19a)$$

$$K_k = P_{k-1}\phi_k\left[\Lambda_k + \phi^{T}_{k}P_{k-1}\phi_k\right]^{-1}, \quad (19b)$$

$$\Lambda_k = \begin{bmatrix} -x^{T}_{k}x_k & 1 \\ 1 & 0 \end{bmatrix}, \quad (19c)$$

$$\phi_k = \begin{bmatrix} w_k & \frac{1}{k}\varphi_k \end{bmatrix}, \quad (19d)$$

$$w_k = \frac{k-1}{k}\left(\hat{\Sigma}^{k-1}_{x\varphi}\right)^{T} x_k, \quad (19e)$$

$$v_k = \frac{1}{k}\begin{bmatrix} (k-1)x^{T}_{k}\hat{\xi}^{k-1}_{xy} \\ y_k \end{bmatrix}, \quad (19f)$$

$$P_k = P_{k-1} - K_k\phi^{T}_{k}P_{k-1} \quad (19g)$$

with $\hat{\Sigma}^{k}_{x\varphi}$ and $\hat{\xi}^{k}_{xy}$ updated as in equations (17d) and (17e), respectively. Since it is the expression P_k which is obtained recursively, hence, in order to calculate $\left(\hat{\Sigma}^{k}_{x\varphi}\right)^{\dagger}$, for the recursive bias compensation equation (16), an additional matrix product has to be computed, i.e.

$$\left(\hat{\Sigma}^{k}_{x\varphi}\right)^{\dagger} = P_k\left(\hat{\Sigma}^{k}_{x\varphi}\right)^{T}. \quad (20)$$

Consequently, the pseudo-inverse of $\hat{\Sigma}^{k}_{x\varphi}$ is obtained in an indirect manner. Moreover, note that the recursive algorithm (19a) requires an inverse of the matrix of dimension 2 × 2 at each recursion. This, however, does not significantly increase the associated computational burden. On the other hand, the important advantages of this algorithm are that, firstly, it can be easily initialised and, secondly, it is relatively insensitive to the quality of the initial values. With reference to (Friedlander, 1984), in the case of no *a priori* information the initialisation can be performed as

$$\Sigma^0_{z\varphi} = \mu \begin{bmatrix} I_{n_\theta} \\ 0_{(n_x - n_\theta)\times n_\theta} \end{bmatrix}, \; P^0_k = \frac{1}{\mu^2} I_{n_\theta}, \; \theta^0_{\mathrm{LS}} = 0_{n_\theta \times 1}. \qquad (21)$$

The scalar parameter μ allows the speed of convergence to be adjusted, hence affects the ‘smoothness’ of $\hat{\Theta}$ (i.e. large value of μ corresponds to the slow convergence and smooth parameters). Further algorithmic details ensuring that the update of the matrix P_k, given by (19g), is (semi-) positive definite are addressed in (Friedlander, 1984).

4.3 Determination of Noise Auto-correlation Elements

Since the matrix W^k is sparse, in general, the computational effort involved in its pseudo-inverse is negligible when compared to that of $\hat{\Sigma}^k_{x\varphi}$. Therefore, it is the pseudo-inverse of $\hat{\Sigma}^k_{x\varphi}$ which forms a crucial bottleneck of the overall algorithm. Consequently, a recursive computation of $\hat{\rho}^k$ is not considered here and its estimate is determined offline at each recursion by solving (14) in the LS sense, i.e.

$$\hat{\rho}^k = \left(W^k\right)^{\dagger} \left(\hat{\xi}^k_{xy} - \hat{\Sigma}^k_{x\varphi} \hat{\theta}^k_{\mathrm{EBCLS}}\right). \qquad (22)$$

5 SIMULATION STUDIES

This section addresses a numerical analysis of the two proposed recursive realisations of the EBPM approach, namely REBPM$_1$ and REBPM$_2$, when applied for the purpose of identifying a SISO discrete-time LTI second order system within the EIV framework. The system to be identified is described by

$$\theta = \begin{bmatrix} -1.5 & 0.7 & 1.0 & 0.5 \end{bmatrix}^T \qquad (23)$$

with the input generated by

$$u_{0_k} = 0.5 u_{0_{k-1}} + \beta_k, \qquad (24)$$

where β_k is a white, zero mean sequence of unity variance. The input noise sequence is zero mean, white of variance $\sigma_{\tilde{u}}$ and the coloured output noise sequence is generated by

$$\tilde{y}_k = 0.7 \tilde{y}_{k-1} + \gamma_k, \qquad (25)$$

where γ_k is zero mean, white and of variance σ_γ. In the case of both algorithms the instrumental vector is based on the instruments proposed in (Ekman et al., 2006), i.e. built from delayed inputs and delayed outputs, and utilised with $n_x = 10$.

Table 1: Results of the estimation of model parameters and auto-correlation elements of the noise sequences.

	true	EBPM	REBPM$_1$	REBPM$_2$
		SNR $\approx$ 11dB		
a_1	-1.500	-1.501 ± 0.041	-1.504 ± 0.051	-1.494 ± 0.023
a_2	0.700	0.701 ± 0.045	0.705 ± 0.056	0.694 ± 0.024
b_1	1.000	0.998 ± 0.039	0.996 ± 0.045	1.001 ± 0.038
b_2	0.500	0.500 ± 0.072	0.495 ± 0.083	0.508 ± 0.051
$\sigma_{\tilde{u}}$	0.100	0.100 ± 0.054	0.095 ± 0.065	0.124 ± 0.052
$r_{\tilde{y}}(0)$	3.922	3.273 ± 2.349	2.647 ± 3.902	3.834 ± 1.376
$r_{\tilde{y}}(1)$	2.745	2.168 ± 1.938	1.631 ± 3.275	2.618 ± 1.174
$r_{\tilde{y}}(2)$	1.922	1.540 ± 0.949	1.250 ± 1.612	1.721 ± 0.715
e_1	–	0.001 ± 0.001	0.004 ± 0.005	0.001 ± 0.001
e_2	–	0.097 ± 0.120	1.187 ± 3.464	0.143 ± 0.197
Λ	–	0	2	0
T	–	–	1.381 ± 0.102	1.663 ± 0.134

The robustness of the two algorithms is examined via a Monte-Carlo simulation study comprising of 100 runs. The mean values of the estimates obtained at the last recursion, i.e. for $k = N$ are recorded and compared with the corresponding results produced by the offline EBPM. The the overall quality of the estimators is assessed via the following two performance criteria:

$$e_1 \triangleq \left\|\hat{\theta}^{\mathrm{N}}_{\lambda} - \theta\right\|^2_2, \qquad e_2 \triangleq \left\|\hat{\rho}^{\mathrm{N}}_{\lambda} - \rho\right\|^2_2, \qquad (26)$$

where λ denotes the λ-th Monte-Carlo run. Prior to the calculation of the performance indeces e_1 and e_2 the possible outliers are removed from the data. An estimate is classified as an outlier if $\left\|\hat{\theta}^{\mathrm{N}}_{\lambda}\right\|_2 > 10$. The number of outliers is denoted by Λ. Additionally, a computation time in seconds, denoted T, is recorded.

The initial values of the parameters are set as follows: $\alpha = 50$ for the REBPM$_1$ and $\mu = 100$ for the REBPM$_2$. In order to provide a fair comparison, in the case of the REBPM$_2$, the bias compensation phase is enabled from sample 50 onwards, although the expressions $\hat{\theta}^k_{\mathrm{LS}}$ and $\left(\hat{\Sigma}^k_{x\varphi}\right)^{\dagger}$ are recursively calculated from the commencement of the algorithm. The values of the noise parameters are chosen as $\sigma_{\tilde{u}} = 0.1$ and $\sigma_\gamma = 2.0$. Consequently, the noise auto-correlation vector is given by

$$\rho = \begin{bmatrix} 3.922 & 2.745 & 1.922 & 0.100 \end{bmatrix}^T, \qquad (27)$$

which yields an approximately equal signal-to-noise ratio (SNR) of around 11dB on both the input and the output signals. The results expressed obtained in terms of mean value ± standard deviation are presented in Table 1. It is observed that the mean values of the model parameters, obtained by the algorithms, see e_1, are relatively accurate and close to the true values and are also characterised by acceptable standard deviations. In the case of e_2 the estimates $\hat{\rho}$ are relatively less precise, especially those produced by the REBPM$_1$.

In general, comparison of the two recursive realisations of the EBPM reveals that it is the REBPM$_2$ which produces the more accurate results overall. Moreover, it is noted that in the case of the REBPM$_1$ the algorithm diverged twice, producing two outliers. In terms of the computational burden, the time required by the REBPM$_2$ is slightly greater when compared to that of REBPM$_1$, i.e. the former technique is faster by approximately 17% with respect to the latter method.

In general, the experiments carried out seem to suggest that the REBPM$_2$ is more advantageous than the REBPM$_1$ due to a simpler initialisation, greater robustness and an absence of convergence problems, at least under the conditions considered here.

6 CONCLUSIONS

A recursive realisation of the extended bilinear parametrisation method for the identification of dynamical linear discrete-time time-invariant single-input single-output errors-in-variables models has been proposed. Two alternative approaches for the online update of the pseudo-inverse of the extended observation covariance matrix have been considered. The first approach is based on the pseudo-inverse matrix lemma, whereas the second is constructed within the framework of the extended instrumental variables technique. For the cases considered, the two resulting algorithms appear to be relatively robust and they are also found to yield precise estimates of the model parameters. Results suggest that the instrumental variables based approach would appear to be the superior of the two developed algorithms.

REFERENCES

Björck, Å. (1996). *Numerical Methods for Least Squares Problems*. SIAM, Philadelphia.

Ekman, M. (2005a). Identification of linear systems with errors in variables using separable nonlinear least squares. In *Proc. of 16th IFAC World Congress*, Prague, Czech Republic.

Ekman, M. (2005b). *Modeling and Control of Bilinear Systems: Applications to the Activated Sludge Process*. PhD thesis, Uppsala University, Sweden.

Ekman, M., Hong, M., and Söderström, T. (2006). A separable nonlinear least-squares approach for identification of linear systems with errors in variables. In *14th IFAC Symp. on System Identification*, Newcastle, Australia.

Feng, D., Zhang, H., Zhang, X., and Bao, Z. (2001). An extended recursive least-squares algorithm. *Signal Proc.*, 81(5):1075–1081.

Friedlander, B. (1984). The overdetermined recursive instrumental variable method. *IEEE Trans. on Automatic Control*, 29(4):353–356.

Ikenoue, M., Kanae, S., Yang, Z., and Wada, K. (2008). Bias-compensation based method for errors-in-variables model identification. In *Proc. of 17th IFAC World Congress*, pages 1360–1365, Seul, South Korea.

Larkowski, T., Linden, J. G., Vinsonneau, B., and Burnham, K. J. (2008). Identification of errors-in-variables systems via extended compensated least squares for the case of coloured output noise. In *The 19th Int. Conf. on Systems Engineering*, pages 71–76, Las Vegas, USA.

Linden, J. G. (2008). *Algorithms for recursive Frisch scheme identification and errors-in-variables filtering*. PhD thesis, Coventry University, UK.

Ljung, L. (1999). *System Identification - Theory for the User*. Prentice Hall PTR, New Jersey, USA, 2nd edition.

Mahata, K. (2007). An improved bias-compensation approach for errors-in-variables model identification. *Automatica*, 43(8):1339–1354.

Markovsky, I. and Van Huffel, S. (2007). Overview of total least-squares methods. *Signal Proc.*, 87(10):2283–2302.

Söderström, T. (2007). Errors-in-variables methods in system identification. *Automatica*, 43(6):939–958.

AN APPROACH OF ROBUST QUADRATIC STABILIZATION OF NONLINEAR POLYNOMIAL SYSTEMS
Application to Turbine Governor Control

M. M. Belhaouane, R. Mtar, H. Belkhiria Ayadi and N. Benhadj Braiek

Laboratoire d'Etude et Commande Automatique de Processus - LECAP
École Polytechnique de Tunis (EPT), BP.743, 2078 La Marsa, Tunis, Tunisie
naceur.benhadj, hela.ayadi, moez.belhaouane@ept.rnu.tn

Keywords: Nonlinear polynomial systems, Lyapunov stability theory, Robust control, Power systems, Linear matrix inequalities.

Abstract: In this paper, robust quadratic stabilization of nonlinear polynomial systems within the frame work of Linear Matrix Inequalities (LMIs) is investigated. The studied systems are composed of a vectoriel polynomial function of state variable, perturbed by an additive nonlinearity which depends discontinuously on both time and state. Our main objective is to show, by employing the Lyapunov stability direct method and the Kronecker product properties, how a polynomial state feedback control law can be formulated to stabilize a nonlinear polynomial systems and, at the same time, maximize the bounds on the perturbation which the system can tolerate without going unstable. The efficiency of the proposed control strategy is illustrated on the Turbine - Governor system.

1 INTRODUCTION

The problem of robust quadratic stabilization for nonlinear uncertain systems has attracted a considerable attention and several methods have been proposed in the literature (Siljak, 1989) (Leitmann, 1993) (Kokotovic and Arcak, 1999). In some very interesting works, Lyapunov stability theory has been used to design control laws for systems with structured or unstructured parametric uncertainties and state perturbations.

The basic principle of quadratic stabilization is to find a feedback controller such that the closed-loop system is stable with a fixed Lyapunov function. This problem was initially proposed in (Barmish, 1985) to study the control of uncertain systems satisfying the so-called matching conditions. Since then, various results have been reported, including a necessary and sufficient condition given in (Barmish, 1985) and the Riccati equation method established in (Petersen and Hollot, 1986). Particularly, they consider the class of linear continuous systems subject to additive perturbations which are nonlinear and discontinuous functions in time and state of the system. The perturbations are uncertain and all we know about them is that they are contained within quadratic bounds.

Recently, the Linear Matrix Inequality (LMI) method has been widely used in quadratic stabilization since it can be solved efficiently using the interior-point method (Boyd et al., 1994). In this fact, the quadratic feedback stabilization of this type of systems using LMI approach has received a great deal of attention in the Siljak-Stipanovic'works (Siljak and Stipanovic, 2000) (Stipanovic and Siljak, 2001) (Siljak and Zecevic, 2005), in which sufficient conditions for quadratic stabilizability are developed. Latter, a new method which gives a less conservative result compared to that of (Siljak and Stipanovic, 2000), by using a descriptor model transformation of the considered system, where an improved sufficient condition for robust quadratic stabilization is given in terms of Linear Matrix Inequality (LMI) (Zuo and Wang, 2005). However, the proposed results remain restrictives to the systems represented by a linear nominal part and these conditions are rather difficult to check and, in general, a nonlinear control law is required.

The main contribution of the present paper consists in the replacement of the linear constant part by a nonlinear polynomial function based on the Kronecker power of the state vector (Rotella and Tanguy, 1988) (Benhadj Braiek et al., 1995) (Brewer, 1978), which has the advantage to approach any analytical nonlinear systems and general enough to model many physical systems (Benhadj Braiek and Rotella, 1992) (Benhadj Braiek and Rotella, 1994) (Benhadj Braiek et al., 1995) (Benhadj Braiek and Rotella, 1995) (Benhadj Braiek, 1995) (Bouzaouache and Benhadj Braiek, 2006). In another hand, we propose the use of the LMI approach in terms of minimization problems (Belkhiria Ayadi and Benhadj Braiek, 2005) , to derive a new sufficient LMI stabilization conditions, which resolution yields a stabilizing polynomial control law involving the quadratic stabilization of the polynomial closed-loop systems and the maximization of the nonlinearity bounds. Notice that, in recent years, various methods have been developed in field of system analysis and control amount to compute the controllers which enlarge the domain of attraction of equilibrium points of polynomial systems through LMI approach (Chesi et al., 1999). Mainly based on LMI relaxation for solving polynomial optimizations (Chesi et al., 2003) (Chesi, 2004), these methods proposes a convex optimization solutions with LMI constraints for a chosen Lyapunov functions (Chesi, 2009).

An additional contribution of this paper is to apply the versatile tools of LMI for the design of robust feedback Turbine-Governor control (Anderson and Fouad, 1977) (Elloumi, 2005). Our primary reason for selecting this type of control is the underlying system model, which can be bounded in a way that conforms to quadratic bounds of nonlinearity. The proposed method has an advantage such as the control design of our power system is formulated as a convex optimization problem, which ensures computational simplicity, and guarantees the existence of a solution. The optimal gains matrices are obtained directly, with no need for tuning parameters or trial and error procedures.

This paper is organized as follows: section 2 is devoted to introduce the description of the nonlinear studied systems and problem formulation. In the third section, the LMI sufficient condition for robust quadratic stabilization of polynomial systems is proposed. Applications to power systems are then considered in section 4. Finally some concluding remarks are given in the last section.

2 DESCRIPTION OF THE STUDIED SYSTEMS AND PROBLEM FORMULATION

In the present paper, we focus on analytical nonlinear polynomial control-affine systems under nonlinear perturbations described by the following state space equation:

$$\begin{aligned}\dot{X} &= f(X(t)) + GU(t) + h(X(t),t) \\ &= \sum_{i=1}^{r} F_i X^{[i]} + GU(t) + h(X(t),t),\end{aligned} \qquad (1)$$

where:
$\forall t \in \mathbb{R}^+$; $X \in \mathbb{R}^n$ is the state vector, $U(t) \in \mathbb{R}^m$ is the input vector.
For $i = 1,...,r$, $X^{[i]} \in \mathbb{R}^{n^i}$ is the *i-th* Kronecker power of the vector X and $F_i \in \mathbb{R}^{n\times n^i}$ are constant matrices.
G is a constant $(n\times m)$ matrix and the polynomial degree r is considered odd: $r = 2s-1$, with $s \in \mathbb{N}^*$.
$h : \mathbb{R}^{n+1} \rightarrow \mathbb{R}^n$ is the nonlinear perturbations. The crucial assumption about nonlinear function $h(t,X(t))$ is that it is uncertain and all we know is that, in the domains of continuity, it satisfies the quadratic inequality:

$$h^T(t,X)h(t,X) \leq \alpha^2 X^T H^T H X, \qquad (2)$$

where $\alpha > 0$ is the bounding parameter and H is a constant matrix. For simplicity, we use $h(t,X)$ instead of $h(t,X(t))$.
A large amount of works have been developed in the robust quadratic stabilization area, considering a particular class of nonlinear systems, where the nonlinearities are totally in the perturbation terms (Stipanovic and Siljak, 2001) (Siljak and Zecevic, 2005) (Zuo and Wang, 2005). The basic idea addressed in this paper is the consideration of a nonlinear uncertain systems described by a polynomial part with nonlinear perturbations which present the uncertainties (Mtar et al., 2007). The present work is an attempt towards expanding the robust quadratic stabilization approach of the considered nonlinear systems in the literature to polynomial ones. The aim of the proposed approach is to guarantees on the one hand, the stabilization of the linear part of the polynomial system, and on the other hand to weaken the perturbation which provide the maximization of the domain of uncertainties.

3 ROBUST STABILIZING CONTROL SYNTHESIS USING THE LMI APPROACH

When the linear part of the perturbed polynomial system (1) (defined by F_1) is not stable, we can introduce a nonlinear feedback to stabilize the overall system and, at the same time, maximize its tolerance to uncertain nonlinear perturbations. The considered polynomial control law is described by the following equation:

$$U = k(X) = \sum_{i=1}^{r} K_i X^{[i]}, \tag{3}$$

where $K_{i,i=1,\dots,r}$ are constant gains matrices, which stabilizes asymptotically and globally the equilibrium $(X = 0)$ of the considered system.
When we apply the feedback (3) to the open-loop system (1), we obtain the closed-loop system:

$$\begin{aligned} \dot{X} &= (f + Gk)(X) + h(t,X) \\ &= a(X) + h(t,X) \\ &= \sum_{i=1}^{r} A_i X^{[i]} + h(t,X) \end{aligned} \tag{4}$$

where:

$$A_i = F_i + GK_i \tag{5}$$

is the closed-loop system matrix.
We define the following set:

$$\mathcal{S}(h,H,\alpha) = \{h : h^T(t,X)h(t,X) \le \alpha^2 X^T H^T H X\}. \tag{6}$$

For any given matrix H, our purpose is to establish robust quadratic stabilization of the system (4-5) and meanwhile make the set $\mathcal{S}(h,H,\alpha)$ as large as possible.

Definition 1. *The system (1) is robustly stabilized by the control law (3) if the closed-loop system (4) is robustly stable with degree α for all $h(t,X)$ satisfying constraint (2).*

Using the quadratic Lyapunov function:

$$V(X) = X^T P X, \tag{7}$$

which is positive definite when P is a symmetric positive definite $(n \times n)$ matrix and computing the derivative $\dot{V}(X)$ along the trajectory of the system (4), lead to the sufficient condition of the global asymptotic stabilization of the perturbed polynomial system. Useful mathematical transformations have allowed the formulation of the obtained condition as an LMI optimization problem according to the polynomial system parameters, given by the following Theorem 1:

Theorem 1. *The system (4) is robustly stabilized by control law (3) if the following optimization problem is feasible:*

$$\textit{minimize} \quad \eta = \frac{1}{\alpha^2}$$

subject to $\mathcal{D}_S(P) > 0,\ \gamma > 0$

$$\begin{bmatrix} \Pi(P) & \star & \star & \star & \star \\ \Lambda\mathcal{D}_S(P)\tau & -I & 0 & 0 & 0 \\ \mathcal{D}_S(P)\tau & 0 & -\frac{1}{\gamma}I & 0 & 0 \\ \mathcal{G}\tilde{\mathcal{M}}(k)\tau & 0 & 0 & -\frac{1}{\gamma}I & 0 \\ H\Lambda\tau & 0 & 0 & 0 & -\eta I \end{bmatrix} < 0 \tag{8}$$

where: $\eta = \frac{1}{\alpha^2}$ *and* $\tilde{\mathcal{M}}(k) = \gamma^{-1}\mathcal{M}(k)$.

$\star$: *denotes the elements below the main diagonal of a symmetric block matrix.*

The relative notations of the Theorem 1 are mentioned in Appendix A.3.
To prove the Theorem 1, we need the two following lemmas:

Lemma 1. *(Yakubovich, 1977)*
Let $\Omega_0(x)$ and $\Omega_1(x)$ be two arbitrary quadratic forms over $\mathbb{R}^n$, then $\Omega_0(x) < 0$for all $x \in \mathbb{R}^n - \{0\}$ satisfying $\Omega_1(x) \le 0$ if and only if there exist $\sigma \ge 0$ such that:

$$\Omega_0(x) - \sigma\Omega_1(x) < 0, \quad \forall x \in \mathbb{R}^n - \{0\} \tag{9}$$

Lemma 2. *(Zhou and Khargonedkar, 1988)*
For any matrices A and B with appropriate dimensions and for any positive scalar $\gamma > 0$, one has:

$$A^T B + B^T A \le \gamma A^T A + \gamma^{-1} B^T B \tag{10}$$

Proof of Theorem 1:
Let us consider the quadratic Lyapunov function (7) and differentiating along trajectory of the system (4), we have:

$$\begin{aligned} \dot{V}(X) &= \dot{X}^T P X + X^T P \dot{X} \\ &= X^T P(\sum_{i=1}^{r} A_i X^{[i]} + h(t,X)) \\ &+ (\sum_{i=1}^{r} A_i X^{[i]} + h(t,X))^T P X \\ &= \sum_{i=1}^{r} (X^T P A_i X^{[i]} + X^{[i]T} A_i^T P X) \\ &+ h^T P X + X^T P h \\ &= 2\sum_{i=1}^{r} X^T P A_i X^{[i]} + h^T P X + X^T P h. \end{aligned} \tag{11}$$

Using the rule of the *vec-function* (see Appendix A.1), one obtains:

$$\dot{V}(X) = 2\sum_{i=1}^{r} \Psi_i^T X^{[i+1]} + h^T P X + X^T P h, \tag{12}$$

where:

$$\Psi_i = vec(PA_i). \tag{13}$$

Then, we have:

$$\begin{aligned} \dot{V}(X) &= 2X^T \mathcal{D}_S(P)\mathcal{M}(a)X + h^T P X + X^T P h \\ &= X^T[\mathcal{D}_S(P)\mathcal{M}(a) + \mathcal{M}(a)^T \mathcal{D}_S(P)]X \\ &+ h^T P X + X^T P h, \end{aligned} \tag{14}$$

where $\mathcal{D}_S(P)$ and $\mathcal{M}(a)$ are defined in Appendix, and $\mathcal{X}$ is expressed by the following equation:

$$\mathcal{X} = \begin{bmatrix} X^T & X^{[2]^T} & \cdots & X^{[s]^T} \end{bmatrix}^T \tag{15}$$

For more details, the transitions between the inequalities (12) to (14) are detailed in (Benhadj Braiek, 1996) (Belhaouane et al., 2008).

When considering the nun-redundant form, the vector $\mathcal{X}$ can be written as:

$$\mathcal{X} = \tau\tilde{\mathcal{X}}, \tag{16}$$

where τ is mentioned in Appendix A.3.

Consequently, $\dot{V}(X)$ can be expressed as:

$$\begin{aligned} \dot{V}(X) &= \tilde{\mathcal{X}}^T\tau^T(\mathcal{D}_S(P)\mathcal{M}(a) + \mathcal{M}(a)^T\mathcal{D}_S(P))\tau\tilde{\mathcal{X}} \\ &+ h^T P X + X^T P h \end{aligned} \tag{17}$$

Which can be written as:

$$\begin{aligned} \dot{V}(X) &= \tilde{\mathcal{X}}^T\tau^T(\mathcal{D}_S(P)\mathcal{M}(a) + \mathcal{M}(a)^T\mathcal{D}_S(P))\tau\tilde{\mathcal{X}} \\ &+ h^T\Lambda\mathcal{D}_S(P)\tau\tilde{\mathcal{X}} + \tilde{\mathcal{X}}^T\tau^T\mathcal{D}_S(P)\Lambda h, \end{aligned} \tag{18}$$

with Λ is mentioned in Appendix A.3.

A sufficient condition of the global quadratic stabilization of the equilibrium ($X = 0$) is that (18) is negative definite. Considering the obtained result, we can derive LMI sufficient conditions for global asymptotic stabilization of the studied system by employing some LMI techniques given by the following development:

Using the S-procedure method, presented by the Lemma1, the inequality (18) with the constraint $h^T h - \alpha^2\tilde{\mathcal{X}}^T(H\Lambda\tau)^T(H\Lambda\tau)\tilde{\mathcal{X}} \leq 0$ derived from (2), is equivalent to the existence of a $\mathcal{D}_S(P)$ matrix and a scalar $\varepsilon \geq 0$ such that:

$$\begin{aligned} &\tilde{\mathcal{X}}^T\tau^T(\mathcal{D}_S(P)\mathcal{M}(a) + \mathcal{M}(a)^T\mathcal{D}_S(P))\tau\tilde{\mathcal{X}} + h^T\Lambda\mathcal{D}_S(P)\tau\tilde{\mathcal{X}} \\ &+\tilde{\mathcal{X}}^T\tau^T\mathcal{D}_S(P)\Lambda h - \varepsilon[h^T h - \alpha^2\tilde{\mathcal{X}}^T(H\Lambda\tau)^T(H\Lambda\tau)\tilde{\mathcal{X}}] < 0 \end{aligned}$$

which can be written as:

$$\begin{bmatrix} \tilde{\mathcal{X}}^T \\ h^T \end{bmatrix}^T \begin{bmatrix} Q_{11} + \varepsilon\alpha^2(H\Lambda\tau)^T(H\Lambda\tau) & \star \\ \Lambda\mathcal{D}_S(P)\tau & -\varepsilon I \end{bmatrix} \begin{bmatrix} \tilde{\mathcal{X}} \\ h \end{bmatrix} < 0 \tag{19}$$

where:
$Q_{11} = \tau^T(\mathcal{D}_S(P)\mathcal{M}(a) + \mathcal{M}(a)^T\mathcal{D}_S(P))\tau$. It should be noted that inequality (19) is a non-strict LMI since $\varepsilon \geq 0$. For the minimization problem, it is well-known in (Boyd et al., 1994), that the minimization result under non-strict LMI constraints is equivalent to that under strict LMI constraints. Thus we can substitute $\varepsilon > 0$ by $\varepsilon \geq 0$. Then, the inequality (19) is further equivalent to the existence of a matrix $\hat{\mathcal{D}}_S(P)$ so that:

$$\hat{\mathcal{D}}_S(P) > 0$$

$$\begin{bmatrix} \hat{Q}_{11} + \alpha^2(H\Lambda\tau)^T(H\Lambda\tau) & \star \\ \Lambda\hat{\mathcal{D}}_S(P)\tau & -I \end{bmatrix} < 0 \tag{20}$$

where $\hat{P} = \varepsilon^{-1}P$ and $\hat{\mathcal{D}}_S(P) = \varepsilon^{-1}D_S(P)$.

In what follows, the matrix $\hat{\mathcal{D}}_S(P)$ is replaced by $\mathcal{D}_S(P)$, to relieve the writing.

According to the following relation (see the lemma given in (Benhadj Braiek et al., 1995)):

$$\mathcal{M}(a) = \mathcal{M}(f + Gk) = \mathcal{M}(f) + \mathcal{G}\mathcal{M}(k), \tag{21}$$

we can write:

$$\begin{bmatrix} S_{11} + R_{11} & \star \\ \Lambda\mathcal{D}_S(P)\tau & -I \end{bmatrix} < 0 \tag{22}$$

where:
$S_{11} = \Pi(P) + \alpha^2(H\Lambda\tau)^T(H\Lambda\tau)$,
$R_{11} = \tau^T(\mathcal{D}_S(P)\mathcal{G}\mathcal{M}(k) + (D_S(P)\mathcal{G}\mathcal{M}(k))^T)\tau$
and $\Pi(P) = \tau^T(\mathcal{D}_S(P)\mathcal{M}(f) + \mathcal{M}(f)^T\mathcal{D}_S(P))\tau$.
$\mathcal{G}$, $\mathcal{D}_S(P)$, $\mathcal{M}(f)$ and $\mathcal{M}(k)$ are mentioned in Appendix A.3.

Using the well known matrix inequality given by Lemma 2, it follows that (22) holds if there exist a constant symmetric matrix $\mathcal{D}_S(P) > 0$ and a positive scalar $\gamma > 0$, such that:

$$\begin{bmatrix} S_{11} + R'_{11} & \star \\ \Lambda\mathcal{D}_S(P)\tau & -I \end{bmatrix} < 0 \tag{23}$$

where:

$$R'_{11} = \gamma\tau^T\mathcal{D}_S(P)^T\mathcal{D}_S(P)\tau + \gamma^{-1}\tau^T\mathcal{M}(k)^T\mathcal{G}^T\mathcal{G}\mathcal{M}(k)\tau_1.$$

Relying on the generalized Schur Complement, equation (23) can be rewritten as:

$$\begin{bmatrix} \Pi(P) + \alpha^2(H\Lambda\tau)^T(H\Lambda\tau) & \star & \star & \star \\ \Lambda\mathcal{D}_S(P)\tau & -I & 0 & 0 \\ \mathcal{D}_S(P)\tau & 0 & -\frac{1}{\gamma}I & 0 \\ \mathcal{G}\mathcal{M}(k)\tau & 0 & 0 & -\gamma I \end{bmatrix} < 0 \tag{24}$$

Pre-multiplying and post-multiplying:

$$\Phi = diag(I, I, I, \gamma^{-1}I)$$

for both sides of (24), we have:

$$\begin{bmatrix} \Pi(P) + \alpha^2(H\Lambda\tau)^T(H\Lambda\tau) & \star & \star & \star \\ \Lambda\mathcal{D}_S(P)\tau & -I & 0 & 0 \\ \mathcal{D}_S(P)\tau & 0 & -\frac{1}{\gamma}I & 0 \\ \mathcal{G}\tilde{\mathcal{M}}(k)\tau & 0 & 0 & -\frac{1}{\gamma}I \end{bmatrix} < 0 \tag{25}$$

where $\tilde{\mathcal{M}}(k) = \gamma^{-1}\mathcal{M}(k)$ and $\tilde{K}_i = \gamma^{-1}K_i$.

Finally, by using the Schur complement, we get:

$$\begin{bmatrix} \Pi(P) & \star & \star & \star & \star \\ \Lambda\mathcal{D}_S(P)\tau & -I & 0 & 0 & 0 \\ \mathcal{D}_S(P)\tau & 0 & -\frac{1}{\gamma}I & 0 & 0 \\ \mathcal{G}\tilde{\mathcal{M}}(k)\tau & 0 & 0 & -\frac{1}{\gamma}I & 0 \\ H\Lambda\tau & 0 & 0 & 0 & -\eta I \end{bmatrix} < 0 \tag{26}$$

where $\eta = \frac{1}{\alpha^2}$.

To establish robust quadratic stabilization in the sense of Definition 1 of the system (4) under the constraint (2) with maximal α, it comes the following optimization problem:

$$\left\{\begin{array}{l} minimize \quad \eta = \frac{1}{\alpha^2} \\ s.t. \quad (26) \\ \gamma > 0, \mathcal{D}_S(P) > 0, P > 0 \end{array}\right. \tag{27}$$

translated by the Theorem1, which ends the proof.

4 APPLICATION TO POWER SYSTEM CONTROL

To illustrate how the proposed LMI approach can be applied for the robust feedback control of power systems, we will consider a electrical mono-machine system with steam valve control (see Figure 1). The parameters of the considered system are given by the following list of symbols:
δ : rotor angle for machine, in radian;
ω : relative speed for machine, in radian/s;
P_m : mechanical power for machine, in pu;
P_c : power control input of machine, in pu;
X_e : steam valve opening for machine, in pu;
H : inertia constant for machine, in second;
D : damping coefficient for machine, in pu;
T_m : time constant of machine's turbine of machine, in second;
T_e : time constant of machine's speed governor, in second;
K_m : gain of machine turbine;
R : regulation constant of machine, in pu;
E : internal transient voltage for machine, in pu;
B : nodal susceptance for machine, in pu;
ω_0 : the synchronous machine speed, in radian/s;
δ_0, P_{m_0} and X_{e_0} are the initial values of $\delta(t)$, $P_m(t)$ and $X_e(t)$ respectively.

The generator dynamics are described as (Sauer and Pai, 1998) (Siljak and Zecevic, 2002):

$$\begin{array}{rcl} \dot{\delta} & = & \omega \\ \dot{\omega} & = & -\frac{D}{2H}\omega + \frac{\omega_0}{2H}(P_m - EVBsin\delta). \end{array} \tag{28}$$

The equation linking the mechanical power P_m to the steam valve opening of turbine X_e for synchronous machine is:

$$\dot{P}_m = -\frac{1}{T_m}P_m + \frac{K_m}{T_m}X_e. \tag{29}$$

The mechanical-hydraulic speed governor can be represented as first order system:

$$\dot{X}_e = -\frac{K_e}{T_e R \omega_0}\omega - \frac{1}{T_e}X_e + \frac{1}{T_e}P_c, \tag{30}$$

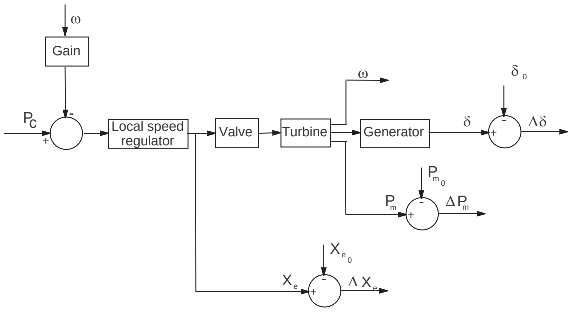


Figure 1: Diagram Bloc representation of Electrical Mono-machine System.

where the term P_c represents the control input. Defining new states:

$$\left[\begin{array}{cccc} \Delta\delta(t) & \omega(t) & \Delta P_m(t) & \Delta X_e(t) \end{array}\right]^T \tag{31}$$

as deviations from the equilibrium,
where:

$$\Delta\delta(t) = \delta(t) - \delta_0 \quad ; \quad \Delta P_m(t) = P_m(t) - P_{m_0}$$
$$\Delta X_e(t) = X_e(t) - X_{e_0}.$$

Then, we obtain the modified system given by the following state space equation:

$$\dot{X}(t) = AX(t) + B_0 U + h(t,X), \tag{32}$$

where:

$$A = \left[\begin{array}{cccc} 0 & 1 & 0 & 0 \\ 0 & \frac{-D}{2H} & \frac{\omega_0}{2H} & 0 \\ 0 & 0 & -\frac{1}{T_m} & \frac{K_m}{T_m} \\ 0 & -\frac{K_e}{T_e R \omega_0} & 0 & -\frac{1}{T_e} \end{array}\right]$$

$B_0 = (0 \quad 0 \quad 0 \quad \frac{1}{T_e})^T$ and

$$h(t,X) = \left[\begin{array}{c} 0 \\ -\omega_0 EVB/2H \\ 0 \\ 0 \end{array}\right] g(X(t)),$$

with $g(X(t)) = sin\delta(t) - sin\delta_0$, represents the nonlinearity of system (32).
The nonlinear system (32) can be developed into a polynomial form by a Taylor series expansions, then we obtain the new state space representation given by the following form:

$$\dot{X} = F_1 X + F_2 X^{[2]} + F_3 X^{[3]} + GU(t) + \bar{h}(t,X), \tag{33}$$

where:

$$F_1 = \left[\begin{array}{cccc} 0 & 1 & 0 & 0 \\ -\omega_0 EVB/2H & \frac{-D}{2H} & \frac{\omega_0}{2H} & 0 \\ 0 & 0 & -\frac{1}{T_m} & \frac{K_m}{T_m} \\ 0 & -\frac{K_e}{T_e R \omega_0} & 0 & -\frac{1}{T_e} \end{array}\right]$$

$$F_2 = \mathbf{0}_{4\times16},\ F_3(2,1) = \frac{\omega_0 EVB}{12H}.$$

For the others values:

$$F_3(i,j) = 0\ \forall i,j : (i = 1,...4; j = 1,...64)$$

$G = B_0$ and

$$\bar{h}(t,X) = \begin{bmatrix} 0 \\ -\omega_0 EVB/2H \\ 0 \\ 0 \end{bmatrix} \bar{g}(X(t)), \qquad (34)$$

where:

$$\bar{g}(X(t)) = sin\delta(t) - \delta(t) + \frac{\delta(t)^3}{6}. \qquad (35)$$

According the values of the machine parameters (Sauer and Pai, 1998) indicated in Table 1, the evolution of the state variables of system (33) is shown in the Figure 2. Since the nonlinearity (34) of system (33) satisfy the quadratic inequality (2) with $H = I$, the LMI robust stabilizing control given by Theorem 1 can be applied to considered power system in order to maximize the domain of nonlinearity while ensuring the stability of the overall system.

Table 1: Machine system parameters.

Symbols of parameters	Values
$H(s)$	5.1
$D(pu)$	3
$T_m(s)$	0.35
$T_e(s)$	0.1
R	0.05
K_m	1
K_e	1
$\omega_0(rad/s)$	314.159

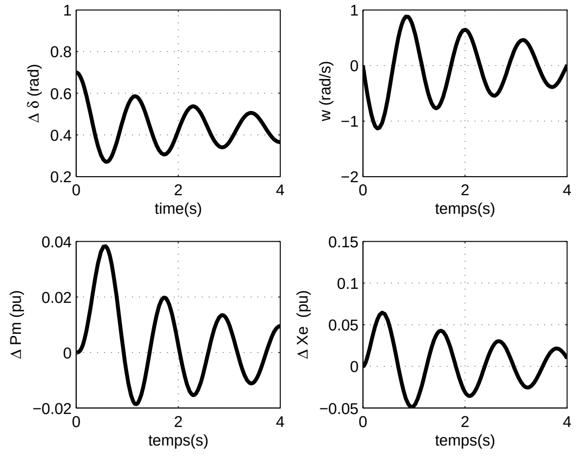


Figure 2: Evolution of state variables towards an perturbation on the variable δ.

The solution of problem (27) yields to the uncertainty bound $\alpha_{max} = 0.4834$ and

$$P = \begin{bmatrix} 10.0734 & 5.3037 & -3.2516 & 1.8013 \\ 5.3037 & 6.0628 & -4.6859 & 1.2175 \\ -3.2516 & -4.6859 & 9.7071 & -2.1715 \\ 1.8013 & 1.2175 & -2.1715 & 12.3111 \end{bmatrix}$$

The control law gain matrices, extracted from $\mathcal{M}(k)$, are given by:

$$K_1 = \begin{bmatrix} -86.592 & -87.171 & -90.911 & -89.173 \end{bmatrix}^T$$

- For $i = 1,...,16$:

$$K_2(1) = -216.909, K_2(2) = K_2(5) = -40.431$$
$$K_2(3) = K_2(4) = K_2(7) = K_2(8) = K_2(9) = -49.443$$
$$K_2(10) = K_2(12) = K_2(13) = K_2(14) = K_2(15) = -49.443$$
$$K_2(11) = K_2(16) = -96.292.$$

- For $i = 1,...,64$:

$$K_3(1) = K_3(17) = K_3(33) = K_3(49) = -216.909$$
$$K_3(2) = K_3(5) = K_3(18) = K_3(21) = -40.431$$
$$K_3(34) = K_3(37) = K_3(50) = K_3(53) = -40.431$$
$$K_3(3) = K_3(4) = K_3(7) = K_3(8) = -49.443$$
$$K_3(9) = K_3(10) = K_3(12) = K_3(13) = K_3(14) = -49.443$$
$$K_3(15) = K_3(19) = K_3(20) = K_3(23) = K_3(24) = -49.443$$
$$K_3(25) = K_3(26) = K_3(28) = K_3(29) = K_3(30) = -49.443$$
$$K_3(31) = K_3(35) = K_3(36) = K_3(39) = K_3(40) = -49.443$$
$$K_3(41) = -49.443 = K_3(42) = K_3(44) = K_3(45) = -49.443$$
$$K_3(46) = K_3(47) = K_3(51) = K_3(52) = K_3(55) = -49.443$$
$$K_3(56) = K_3(57) = K_3(58) = K_3(60) = K_3(61) = -49.443$$
$$K_3(62) = K_3(63) = -49.443$$
$$K_3(6) = K_3(11) = K_3(16) = K_3(22) = K_3(27) = -96.292$$
$$K_3(32) = K_3(38) = K_3(43) = K_3(54) = K_3(59) = -96.292$$
$$K_3(64) = -96.292$$

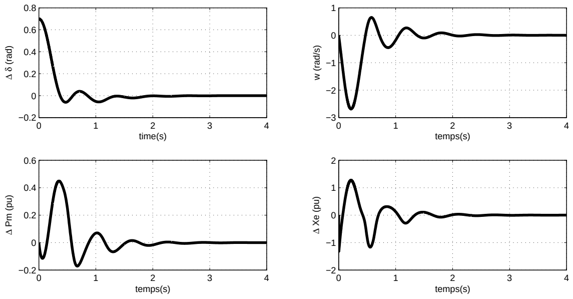


Figure 3: Closed-loop responses of the power system with polynomial control.

From the simulation results shown in Figure 3, it is obvious that the results confirm the validity of the proposed method and the uncertainty bound found by the LMI procedure (27), dominates the maximum of the perturbation function (35) of the system (33). The polynomial robust control can rapidly damp the oscillations of the studied system and greatly enhance transient stability of the mono-machine power system. Besides, the polynomial control is more reassuring in the case of a more aggressive perturbation.

5 CONCLUSIONS

A sufficient LMI condition for robust quadratic stabilization of polynomial systems under nonlinear perturbations has been proposed in this work. This new feedback stabilizing approach is based on the direct Lyapunov method and elaborated algebraic developments using the Kronecker product properties. These developments have been turned into an LMI minimization problem, which can be easily solved by means of numerically efficient convex programming algorithms. A mono-machine power system is considered as an application example of the technique developed in this paper. The numerical simulation results have confirmed the efficiency of the proposed polynomial controller which can rapidly damp the system oscillations and greatly enhance the transient stability of the considered mono-machine power system despite the nonlinear uncertainty affecting the studied system.

REFERENCES

Anderson, P. M. and Fouad, A. A. (1977). Power system control and stability. *The IOWA*.

Barmish, B. (1985). Necessary and sufficient conditions for quadratic stabilizability of an uncertain systems. *Journal of Optimization Theory and Applications*, 46:399–408.

Belhaouane, M., Mtar, R., Belkhiria Ayadi, H., and Benhadj Braiek, N. (2008). An LMI technique for the global stabilization of nonlinear polynomial systems. *International Journal of Computers, Communication and Control (IJCCC), to appear*.

Belkhiria Ayadi, H. and Benhadj Braiek, N. (2005). On the robust stability analysis of uncertain polynomial systems: an LMI approach. *17th IMACS World Congress, Scientific Computation, Applied Mathematics and Simulation*.

Benhadj Braiek, N. (1995). Feedback stabilization and stability domain estimation of nonlinear systems. *Journal of The Franklin Institute*, 332(2):183–193.

Benhadj Braiek, N. (1996). On the global stability of nonlinear polynomial systems. *IEEE Conference On Decision and Control,CDC'96*.

Benhadj Braiek, N. and Rotella, F. (1992). Robot model simplification by means of an identtification method. *Robotics and Flexible Manufacturing Systems, Edit. J. C. Gentina and S. G. Tzafesta*, pages 217–227.

Benhadj Braiek, N. and Rotella, F. (1994). State observer design for analytical nonlinear systems. *IEEE Syst. Man and Cybernetics Conference*, 3:2045–2050.

Benhadj Braiek, N. and Rotella, F. (1995). Stabilization of nonlinear systems using a Kronecker product approach. *European Control Conference ECC'95*, pages 2304–2309.

Benhadj Braiek, N., Rotella, F., and Benrejeb, M. (1995). Algebraic criteria for global stability analysis of nonlinear systems. *Journal of Systems Analysis Modelling and Simulation, Gordon and Breach Science Publishers*, 17:221–227.

Bouzaouache, H. and Benhadj Braiek, N. (2006). On guaranteed global exponential stability of polynomial singularly perturbed control systems. *International Journal of Computers, Communications and Control (IJCCC)*, 1(4):21–34.

Boyd, S., Ghaoui, L., and Balakrishnan, F. (1994). Linear Matrix Inequalities in System and Control Theory. *SIAM*.

Brewer, J. (1978). Kronecker product and matrix calculus in system theory. *IEEE Trans.Circ.Sys*, CAS-25:722–781.

Chesi, G. (2004). Estimating the domain of attraction for uncertain polynomial systems. *Automatica*, 40(11):1981–1986.

Chesi, G. (2009). Estimating the domain of attraction for non-polynomial systems via LMI optimizations. *Automatica, doi:10.1016/j.automatica.2009.02.011(in press)*.

Chesi, G., Garulli, A., Tesi, A., and Vicino, A. (2003). Solving quadratic distance problems: an lmi-based approach. *IEEE Transactions on Automatic Control*, 48(2):200–212.

Chesi, G., Tesi, A., Vicino, A., and Genesio, R. (1999). On convexification of some minimum distance problems. *5th European Control Conference ECC'99*.

Elloumi, S. (2005). *Commande décentralisée robuste des systèmes non linéaires interconnectés: Application à un système de production-transport de l'énergie éléctrique multi-machines*. Thèse de doctorat es sciences, Ecole Nationale d'Ingénieurs de Tunis.

Kokotovic, P. and Arcak, M. (1999). Constructive nonlinear control: Progress in the 90's. *Proceedings of the 14th IFAC World Congress, Beijing, P.R. China*, Plenary Volume:49–77.

Leitmann, G. (1993). On one approach to the control of uncertain systems. *Journal of Dynamic Systems, Measurement and Control*, 115:373–380.

Mtar, R., Belkhiria Ayadi, H., and Benhadj Braiek, N. (2007). Robust stability analysis of polynomial systems under nonlinear perturbations: an LMI approach. *Fourth Inernational Multi-conference on Systems, Signals and Devices (SSD'07)*.

Petersen, I. and Hollot, C. (1986). A Riccati equation approach to the stabilization of uncertain linear systems. *Automatica*, 22:397–411.

Rotella, F. and Tanguy, G. (1988). Non linear systems: identification and optimal control. *Int.J.Control*, 48(2):525–544.

Sauer, P. W. and Pai, M. A. (1998). Power system dynamics and stability. *Englewood Cliffs, NJ:Prentice-Hall*.

Siljak, D. (1989). Parameter space methods for robust control design: A guided tour. *IEEE Transactions on Automatic Control*, 34:674–688.

Siljak, D. and Stipanovic, D. (2000). Robust stabilization of nonlinear systems: The LMI approach. *Mathematical Problems in Engineering*, 6:461–493.

Siljak, D. and Zecevic, A. (2005). Control of large-scale systems: Beyond decentralized feedback. *Annual Reviews in Control*, 29:169–179.

Siljak, D. and Zecevic, D. M. S. A. (2002). Robust decentralized turbine/governor control using linear matrix inequalities. *IEEE Transactions On Power Systems*, 17(3):715–722.

Stipanovic, D. and Siljak, D. (2001). Robust stability and stabilization of discrete-time nonlinear systems: The LMI approach. *International Journal of Control*, 74:873–879.

Yakubovich, V. (1977). S-procedure in nonlinear control theory. *Vestnick Leningrad Univ. Math.,*, 4:73–93.

Zhou and Khargonedkar (1988). Robust stabilization of linear systems with norm-bounded time-varying uncertainty. *Sys. Contr. Letters*, 10:17–20.

Zuo, Z. and Wang, Y. (2005). A descriptor system approach to robust quadratic stability and stabilization of nonlinear systems. *IEEE International Conference on Systems, Man and Cybernetics*, 1:486–491.

APPENDIX

The dimensions of the matrices used in this section are the following: $A(p\times q), C(q\times f), E(n\times p)$

A.1- *vec(.)* function:

An important vector valued function of matrix denoted $vec(.)$ was defined in (Brewer, 1978) as follows:

$$A = \left[\begin{array}{cccc} A_1 & A_2 & \dots & A_q \end{array}\right] \in \mathbb{R}^{p\times q},$$

where

$$\forall i \in \{1,...,q\},\, A_i \in \mathbb{R}^p$$
$$vec(A) = [A_1 \;\; A_2 \;\; \dots \;\; A_q]^T \in \mathbb{R}^{pq}.$$

We recall the following useful rule (Brewer, 1978) of *vec-function*:

$$vec(EAC) = (C^T \otimes E)vec(A)$$

A.2- *mat*(.) function:

A special function $mat_{(n,m)}(.)$ can be defined as follows:

If V is a vector of dimension $p = n.m$ then $M = mat_{(n,m)}(V)$ is the $(n\times m)$ matrix verifying $V = vec(M)$.

A.3- Notations related to Theorem 1:

(i).

$$\tau = \begin{bmatrix} T_1 & & & & \\ & T_2 & & 0 & \\ & & T_3 & & \\ & 0 & & \ddots & \\ & & & & T_s \end{bmatrix}$$

where:

$$\forall i \in \mathbb{N};\ \exists!\ T_i \in \mathbb{R}^{n^i\times n_i},\ such\ as\ X^{[i]} = T_i\tilde{X}^{[i]},$$

with:

$\tilde{X}^{[i]}$ represents the nun-redundant i-power of X (Benhadj Braiek et al., 1995).

(ii).

$$\Lambda = \left[\begin{array}{cccc} I_n & 0_{n\times n^2} & \cdots & 0_{n\times n^s} \end{array}\right],$$

verifying $X = \tilde{X} = \tilde{\Lambda}\tilde{X},\ \Lambda^T P = \mathcal{D}_s(P)\Lambda^T$ where:

$$\tilde{\Lambda} = \Lambda\tau.$$

(iii).

$$\mathcal{D}_s(P) = \begin{bmatrix} P & & & 0 \\ & P\otimes I_n & & \\ & & \ddots & \\ 0 & & & P\otimes I_{n^{s-1}} \end{bmatrix}$$

(iv).

$$\mathcal{G} = \begin{bmatrix} G & & & 0 \\ & G\otimes I_n & & \\ & & \ddots & \\ 0 & & & G\otimes I_{n^{s-1}} \end{bmatrix}$$

(v).

$$\Pi(P) = \mathcal{D}_S(P)\mathcal{M}(f) + \mathcal{M}(f)^T\mathcal{D}_S(P),$$

where for a polynomial vectorial function:

$$z(X) = \sum_{i=1}^{r} Z_i X^{[i]},$$

with $X \in \mathbb{R}^n$ and Z_i are $(n\times n^i)$ constant matrices. We define the $(\upsilon\times\upsilon)$ matrix $\mathcal{M}(z)$ by:

$$\mathcal{M}(z) = \begin{bmatrix} M_{11}(Z_1) & M_{12}(Z_2) & 0 & \dots & 0 \\ 0 & M_{22}(Z_3) & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & M_{s-1,s-1}(Z_{2s-3}) & M_{s-1,s}(Z_{2s-2}) \\ 0 & \dots & \dots & 0 & M_{s,s}(Z_{2s-1}) \end{bmatrix},$$

where $\upsilon = n + n^2 + ... + n^s$ and:

- For $j = 1,...,s$

$$M_{j,j}(Z_{2j-1}) = \begin{bmatrix} mat_{(n^{j-1},n^j)}\left(Z_{2j-1}^{1T}\right) \\ mat_{(n^{j-1},n^j)}\left(Z_{2j-1}^{2T}\right) \\ \vdots \\ mat_{(n^{j-1},n^j)}\left(Z_{2j-1}^{nT}\right) \end{bmatrix}$$

- For $j = 1,...,s-1$

$$M_{j,j+1}(Z_{2j}) = \begin{bmatrix} mat_{(n^{j-1},n^j)}\left(Z_{2j}^{1T}\right) \\ mat_{(n^{j-1},n^j)}\left(Z_{2j}^{2T}\right) \\ \vdots \\ mat_{(n^{j-1},n^j)}\left(Z_{2j}^{nT}\right) \end{bmatrix}$$

where Z_k^i is the i^{th} row of the matrix Z_k.

BUILDING TRAINING PATTERNS FOR MODELLING *MR* DAMPERS

Jorge de-J. Lozoya-Santos, Javier A. Ruiz-Cabrera, Vicente Diaz-Salas
Ruben Morales-Menendez and Ricardo Ramirez-Mendoza
Tecnológico de Monterrey, Av. Garza Sada 2501, Monterrey, NL, México
{jorge.lozoya, A00804994, A00790521, rmm, ricardo.ramirez}@itesm.mx

Keywords: Pattern validation, Modelling, Magneto-rheological damper, Identification.

Abstract: A method for training patterns validation for modelling *MR* damper is proposed. The method was validated with two models based on black-box and semi-phenomenological approaches. An input pattern that allows a better identification of the *MR* damper model was found. Including a frequency modulated displacement and increased clock period in the training pattern, the *MR* damper model fitting is improved. Also, the designed input pattern minimizes of training phase and reduces of the number of experiments. Additionally, incorporation of the electric current in the *MR* models outperforms the modelling approach.

1 INTRODUCTION

A Magneto-Rheological (*MR*) damper is a device which allows the dissipation of energy in a automotive suspension. Its principal components are: piston, housing, accumulator, coil and *MR* fluid. The mechanical structure is the unchanged of passive damper. The *MR* fluid is a suspension of micrometer-sized magnetic particles in an oil. In the area of piston where the oil is transfered between housing chambers, a magnetic field is applied via electric current. The later varies the damping properties of the device. The interaction of the several aforementioned mechanisms and the magnetic field variations results in highly non linear behavior with hysteretic patterns of the generated force.

The power generated and energy dissipated by this device are defined by the piston displacement and velocity multiplied for the force, respectively. In the control system of a semi-active automotive suspension, the precision on the generated power and dissipated energy of *MR* damper is crucial. This precision depends on the accuracy of the *MR* damper model. Therefore, a good *MR* damper model is a key issue. Having a good *MR* damper simulation demands a good mathematical equation of the damper, a good training phase and a good learning phase of the model (coefficients). The training phase must find out the main characteristics of the *MR* damper through the training patterns. This requires a specific Design of Experiments (*DoE*).

The role that the input patterns plays in the *MR* damper identification process was analyzed with two models. The hypothesis is that there exists experimental input patterns that allows the best learning of the coefficients of the model, regardless the chosen model structure. This paper is organized as follows. Section 2 presents a literature review. In section 3, several input patterns were implemented in order to validate the proposal. Section 4 discusses the results. Section 5 concludes the paper.

2 LITERATURE REVIEW

2.1 Input Patterns

The training patterns of the most representative modeling approaches were reviewed. In some research works the Design of Experiments (*DoE*) and the *MR* damper model were not clearly associated. The development of a *MR* damper model, its parameterization and its final application were not integrally performed.

A training pattern for modelling of a *MR* damper consists of signal that includes a displacement (x) and electric current (I). The velocity is considered as a rate of change of the displacement. Table 1 summarizes some important works. This table is divided in three sections according to the training patterns.

Section one (SSS+C). The displacement is a

Table 1: Comparison of *MR* damper models. *P*ower column shows if the research work studied the generated power (✓), or did not (✗). *E*nergy column same meaning as power column. Finally, the reference are cited.

Power	Energy	SSS+C training patterns
✓	✓	(Spencer et al., 1996)
✓	✓	(Li et al., 2000)
✓	✗	(Wang and Kamath, 2006)
✗	✗	(Shivaram and Gangadharan, 2007)
✓	✓	(Guo et al., 2006)
✓	✓	(Nino-Juarez et al., 2008)
Power	Energy	BWN+C training patterns
✗	✓	(Burton et al., 1996)
✗	✗	(Wang and Liao, 2001)
✗	✗	(Savaresi et al., 2005)
Power	Energy	BWN+BWN training patterns
✗	✗	(Wang and Liao, 2001)
✗	✗	(Chang and Zhou, 2002)
✓	✓	(Du et al., 2006)

*S*inusoidal *S*weep *S*ignal (*SSS*) with specific frequency and a *C*onstant electric current. This is a typical training input for *MR* damper models. Exploiting this pattern, both *E*nergy (*E*) and *P*ower (*P*) are successfully simulated. The number of experiments is high. The obtained model accuracy is high (5% error). There are not electrical current transients, this compromises the use of the model. Table 1 shows six representative works with this type of training inputs.

Section two (BWN+C). The displacement is a bandwidth *B*and*W*idth *N*oise (*BWN*) pattern and a *C*onstant electric current (*C*). This training input patterns has the same features as SSS+C signals. However, the information richness due to the magnitude of displacement is decreased. *P*ower and *E*nergy simulation are not achieved.

Section three (BWN+BWN), both displacement and electric current follow a bandwidth noise *BWN* pattern. Power and energy simulation are not achieved, except if a displacement greater than 10 mm is generated, (Du et al., 2006). The pattern requires shorter training inputs. The fitting error this pattern is low (3%).

The *B*and*W*idth (*BW*) in all the reviewed works was lower than 6 Hz, except in (Savaresi et al., 2005) and (Nino-Juarez et al., 2008). Hence, the *MR* damper response to high frequency has not been explored. Neither, hard non-linearities due to the broken magnetic bounds of metallic particles (because of high frequency and displacements). There are missing analysis of power and energy responses in automotive applications for frequencies around 10-15 Hz and displacement greater than 10 mm.

There are research works with other type of patterns, such as, *A*mplitude *P*seudo *R*andom *B*inary *S*ignal *APRBS* in electric current (Savaresi et al., 2005). (Wang and Liao, 2005). explored electric current with sinusoidal wave signals.

There are not a standard definition of training patterns in order to identify the power and energy features. The overuse of the *MR* damper due to long experimental exploration at electric current greater than 3 amperes could give a skewed model. Therefore, more research of training patterns for *MR* damper modelling is needed.

2.2 Modeling Approaches

Several models have been developed with different approaches. These models could be: phenomenological (*P*), semi-phenomenological (*SP*) and black-box (*BB*) (neural network, fuzzy, non-linear ARX, polynomial among others). The training pattern will be tested with both non-linear with *A*uto *R*egressive e*X*ogenous inputs (*NARX*) and a *S*emi-*P*henomenological models. A brief review of these models will be included for completeness.

Table 2: Description of variables for DoE.

Variable	Description
x_k or x	Damper piston displacement
I_k or I	Electrical current
$\dot{x}_k$ or $\dot{x}$	Damper piston velocity
f_{MRk} or f_{MR}	Damping force
a_j	j-esime modeling coefficient
d	time delays
q_1, q_2	Electrical current exponents
ESR	Error-Signal-to-noise Ratio
k	Discrete sample, discrete time
j	Subindex

The *MR* damper model based on a non-linear ARX structure is a lineal combination of a vector of delayed inputs multiplied for their parameters. If electric current is not an input, all the parameters have a polynomial dependence on it.

In (Nino-Juarez et al., 2008), a non-linear ARX model of nine parameters (1) achieves high precision simulation of power and energy. Table 2 defines the parameters of this equation.

$$\begin{aligned} f_{MRk} &= a_1 f_{MRk-1} + a_2 f_{MRk-2} + a_3 f_{MRk-3} \\ &\quad + a_4 x_{k-1} + a_5 x_{k-2} + a_6 x_{k-3} \\ &\quad + a_7 \dot{x}_{k-1} + a_8 \dot{x}_{k-2} + a_9 \dot{x}_{k-3} \end{aligned} \tag{1}$$

By the side of Semi-Phenomenological (*SP*) approaches, the bi-viscous and hysteretic behavior are shaped with smooth and concise forms. The instantaneous force is delivered without taking into account the transients, and consequently at high frequency, dynamic features are not well emulated. Transients can be omitted at both low frequencies and small displacements. The coefficients are related to energy and power features of *MR* dampers but they are not linked to components. The SP model has a good balance between simulation capability and easy to fit model.

(Guo et al., 2006) (2) have well defined parameters for the dynamic yield force, the post-yield and the pre-yield proportions. The the *MR* damper response is simulated using hyperbolic tangents.

$$f_{MR} = a_1 tanh\left(a_3\left(\dot{x} + \frac{a_4}{a_5}x\right)\right) + a_2\left(\dot{x} + \frac{a_4}{a_5}x\right) \tag{2}$$

The non-linear ARX model has less than 1% error prediction; while the SP approach has less than 4% error prediction.

3 EXPERIMENTS

The specimen tested was a DELPHI Gabriel *MR* damper. It is a standard mono-tube configuration with 36 mm piston and *MR* fluid. This damper is part of the Delphi MagneRide^{TM} commercial system. The configuration of the experimental system was a MTS^{TM} which can deliver enough force and time response with respect to the maximum force and bandwidth of the *MR* damper.

The monitored variables were the damping force f_{MR}, displacement x, and electric current I. The data acquisition system was Software Testlink^{TM} and Testware^{TM} II). Thanks to *Metalsa*[1] for using its facility.

The DoE considered a displacement that follows this signal $0.0125 \cdot sin(\omega \cdot \frac{k}{512})$, where the sampling frequency was 512Hz, $\omega = 2\pi f$, $f = \{1, 1.5, \ldots, 13.5, 14\}$ Hz. The absolute resultant range for the displacement was $[0, 25]$ mm. The absolute resultant range for the force was $(0, 2.850]$ N. The current was kept constant.

The displacement signal was replicated 12 times. At each replicate, the current was increased in 0.25 A from 0 to 4 A. Figure 1 shows an example of these experimental results.

All the experimental data sets were identified with a semi-phenomenological model obtaining a *MR* damper simulator (Guo et al., 2006).

In order to generate several datasets, eleven training patterns were designed. The Table 3 shows two example of this. The displacement follows a *F*requency *M*odulated *FM* signal with fixed amplitude and a *BW* from 0.5 to 14.5 Hz. This signal was generated by a Voltage Controlled Oscillator (*VCO*). The *VCO*'s input was an *ICPS* with values between 0 and 1. Each ICPS step had a time length of 100*ms* which means that the frequencies in signal were constant over same time. The magnitude of displacement was held constant in 3 mm. Two different signals of electric current were evaluated: an *I*ncreased *C*lock *P*eriod *S*ignal (*ICPS*) and a *P*seudo *R*andom *B*inary *S*equel (*PRBS*). The length of time was 30 seconds. All the experiments were piecewise designed, assuring the invariance of conditions during all the experiment. The signals were fed through the simulator in order to retrieve the force.

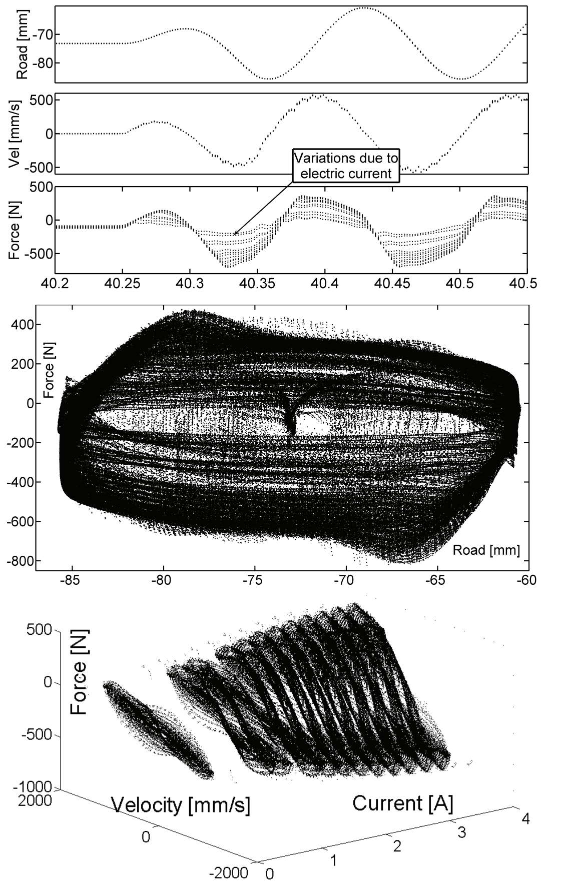


Figure 1: Experimental data for 4.5 – 14.5 Hz (high frequency). Top plot. Time versus displacement, velocity and force response. Middle and bottom plots show the experimental energy and power. These plots include several force responses according to applied electric currents.

For the displacement, the several DoE forms were: *S*inusoidal *S*tepped *F*requency (*SFS*), *S*inusoidal *CH*irp Signal (*CHS*), *R*oad *P*rofile (*RP*) and (*FM*).

For the electric current, the DoE shapes were:

[1] www.metalsa.com.mx

*S*tepped increment at electrical *C*urrent (*SC*), *R*amp Periodic positive slope *C*urrent (*RC*), *R*amp with positive and negative slope *C*urrent (*ADRC*), *ICPS*, and *PRBS*.

With later defined signals the DoEs were:(1) (SFS,SC), (2)(SFS,RC), (3)(SFS,ADRC), (4)(CHS,ICPS), (5)(CHS,PRBS), (6)(SFS,ICPS), (7)(SFS,PRBS), (8)(FM,ICPS), (9)(FM,PRBS), (10)(RP,ICPS) and, (11)(RP,PRBS). The precedent number will identified the training patterns in the rest of the paper. For rich details, see (Lozoya-Santos et al., 2009).

Table 3: Design of experiments.

PTIC	Displacement		Current	$\frac{PTIC(I)}{PTIC(x)}$
	$\tau(vco)$	*Time*		
FM,ICPS	0.10s	30	$\tau_{\|I\|} = 0.10s$	1
FM,PRBS	0.10s	30	$min_\tau = 0.05s$	1

Equation	Description
$\tau(vco)$	Time constant for VCO
$\frac{PTIC(I)}{PTIC(x)}$	How many PTIC(I) utilized each PTIC(x)
min_τ	Minimum clock period in PRBS
$\tau_{\|I\|}$	Amplitude Period in ICPS

The Figure 2 shows three computed experiments. These training pattern exhibit fixed amplitude for displacement and persitent signals in frequency.

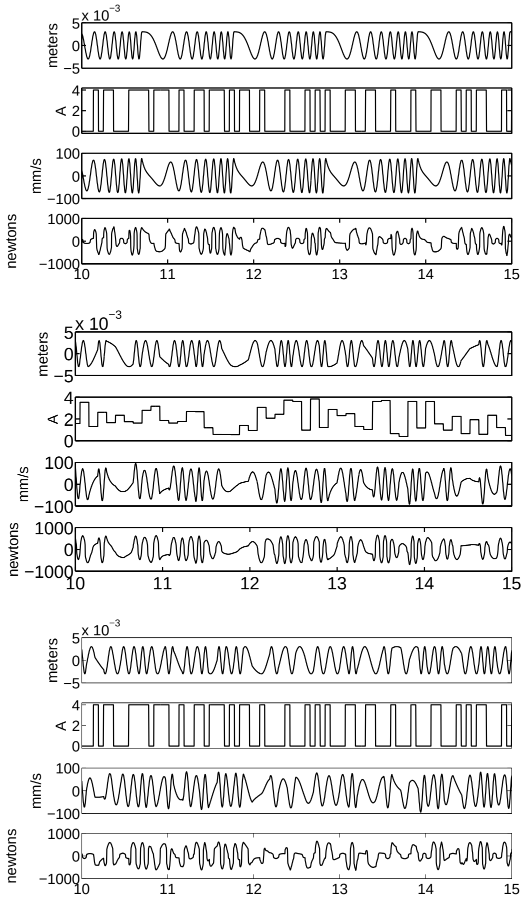


Figure 2: Experimental training patterns. Top plot shows a (*CHS,PRBS*). Middle plot shows a (*FM,ICPS*). Bottom plot describes a (*FM,PRBS*).

4 RESULTS

Performing an analysis of the models (1) and (2) and a-priori knowledge of *MR* damper dynamics, the modified models and its degree of freedoms were proposed. The *D*egree *o*f *F*reedom (*DoF*) of the model represents a main variation in the original structure of the model

Then, for the given experimental data sets, the two *MR* damper models were trained. The first trained model was the modified version of (1). The new model has added regressors of the current. Each delayed value of current is raised to power two. Thus, the augmented regressors structure were:

$$a_{10} \cdot I_{k-1}^2 + \cdots + a_{\{10+d-1\}} \cdot I_{k-(1+d)}^2 \qquad (3)$$

The resultant model could have from 11 to 14 parameters, then its *DoF* is the number of parameters for I. The second trained model was the modified SP model (4). The modification consisted of the incorporation of the factor I^{q_j} as direct input on both terms, where j={1,2} is the j-esime model term. The *DoF*s of model (4) were the power q_j, where its possible values were {0.5, 0.33, 0.25, 0.2}. The parameters number remains the same.

$$f_{MR} = a_1 \cdot I^{q_1} tanh\left(a_3\left(\dot{x} + \frac{a_4}{a_5}x\right)\right) + a_2 \cdot I^{q_2}\left(\dot{x} + \frac{a_4}{a_5}x\right) \qquad (4)$$

Finally, the modified models were identified by nonlinear curve fitting using the non-linear least squares algorithm. Based on these models, the DoEs were validated.

The training patterns have persistent signals with richness frequency content for each signal (x, $\dot{x}$, I, f_{MR}). For each experiment, the model was fitted at each defined *DoF*. Then a validation of the model via ESR was performed with the rest of the experiments and the ESR average was computed.

This process was repeated until all the possible values of *DoF* were varied and the resultant model was fitted.

After obtaining all the fit measures per DoF for experiment data sets, a sort process from lowest to highest ESR was done. This step did include all the experiments. The combination of *DoF* and experiment with the lowest error-to-noise ratio was selected as the best.

Identification and validation were performed for all the experiments and models. Therefore proposed NARX model was submitted to the variation in number of coefficients. The semi-phenomenological model always maintained five coefficients along *DoF* variations. The models include the electric current as natural input. The validation process confirmed that the emulation of bi-viscous and hysteretic features by the proposed models are dependent on the design of experiments.

Table 4: Comparison of ESR results for different training patterns. M is the type of model. E is the number of training pattern, ESR AVG is the average ESR. BEST ESR is the best obtained ESR.

M	E	ESR AVG	BEST ESR
BB	8	0.002	0.0009
	9	0.003	0.0009
	5	0.0011	0.0010
	11	2139	2232
	6	2.0582	3.6940
	7	1.1010	1.9761
SP	8	0.0239	0.0212
	1	0.0235	0.0202
	4	0.0240	0.0205
	10	0.1916	0.1055
	3	0.1002	0.1094
	11	0.2132	0.0979

In Table 4, general approach results are shown. BB and SP correspond to the modified model (1) and the model (4) respectively. E column sorts the experiments by the top 3 performance and the worst 3 for the same DoF. The next columns ESR AVG and SD shows the overall ESR statistical behavior, in other words, for all DoF variations and for all validations with a specific experiment. For example, the first row specifies that EXP 8 for NARX proposed model has an average $ESR = 0.0002$ when the coefficients obtained by experiment 8 are used to validate others patterns. For this row, the best model has a $DoF = 2$. An analysis of the full table demonstrates that the best performing E is the number 8 because it has the lowest ESR values. For completeness, the values for DoFs have been included into the Table 4. The best performing *DoF* were: number of regressors of electric current equal to 2 for NARX model and $q_1 = 0.5, q_2 = 0.2$ for electric current dependent semi-phenomenological model.

By other side, the experiments 11, 6 and 7 used to fit BB and the experiments 10, 3 and 11 used to fit SP have big ESRs (i. e. lack of fit). The experiment 11 is repeated for both worst cases, hence the use of road profiles could generate skewed models. Thus, the configuration of input patters has high signification on the learning of model parameters.

Based on the results, a frequency modulated displacement, with the same spectral frequency content as road profile and an electrical current excitation with ICPS shaping can recreate the dynamical force response of *MR* Damper devices, regardless of the *MR* damper model's structure.

Moreover, coefficients are robust when model is tested with other patterns (cross validation), obtaining lows ESRs. The ESR span intervals were for NARXs $(6.17x10^{-5}, 0.00026)$, for SP $(0.00635, 0.05687)$ and for P $(0.02234, 0.06917)$, respectively. The experiments 5, 7, 9 y 12 (current equal to PRBS) for modified SP always offer an $ESR \geq 0.15$ which means that discontinuous values of current are not proper for model.

4.1 Discussion

The classic DoE has poor frequency content in electric current and excessive repetitions in displacement. Hence, the number of experiments are a multiple of the values of the tested electric current plus the replicates for each experiment. The lengths of time of the eleven DoEs in this work were between 30-60 seconds. The maximum number of experiments will be 30, including the replicates. The realistic nature of exogenous and actuation variables allows a safe test.

The best learning of the coefficients in each tested model was successfully with the training pattern *FM+ICPS*. The frequency modulated displacement implies a continuous changes of slope implying the persistence of the effect of the velocity over *MR* damper. Therefore, with a short and continuous test, the uniform coverage of the semi-active zone, (the exploration of energy and power capabilities) in *MR* damper is achieved.

5 CONCLUSIONS

A comparative analysis of training pattern for identification of *MR* damper models was done. Two models were exploited to validate the proposal: non-linear ARX and Semi-phenomenological models. The key variables in training patterns are the frequency bandwidth and the electric current.

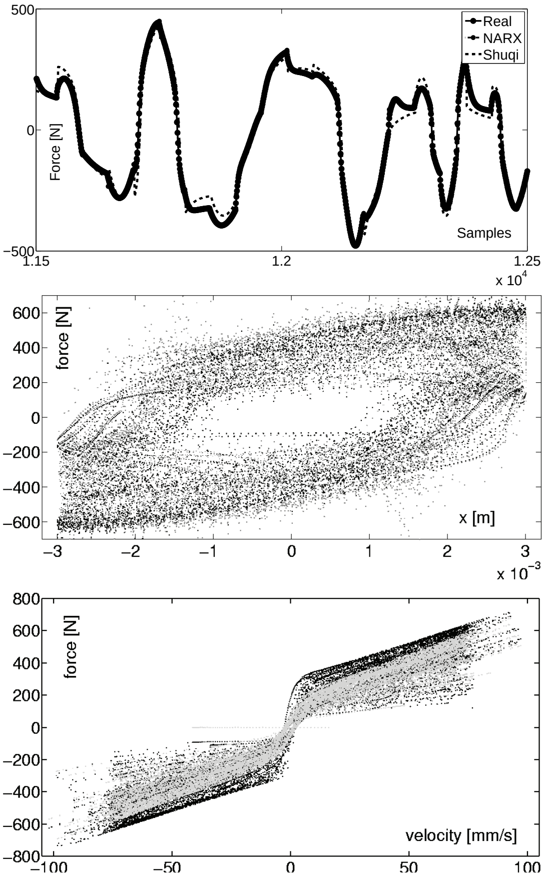


Figure 3: Top plot. Comparison of simulated responses versus experiment 8. Middle and bottom plots show the power and energy extraction using the experiment 8. The black dots are experimental data. The gray dots are emulated data by SP proposed model with a $q_1 = 0.5$, $q_2 = 0.2$.

I was validated that the configuration of the training patterns has a high impact over the model fitting. Features as short duration, continuity, uniform coverage of electric current and displacement ranges are needed. Also, the use of the model must be considered into the DoE.

REFERENCES

Burton, S., Makris, N., Konstantopoulos, I., and Antsaklis, P. (1996). Modeling the Response of ER Damper: Phenomenology and Emulation. *Eng. Mech.*, 122:897–906.

Chang, C. and Zhou, L. (2002). Neural Network Emulation of Inverse Dynamics for a MR Damper. *Struct. Eng.*, 128:231–239.

Du, H., Lam, J., and Zhang, N. (2006). Modelling of a Magneto-Rheological Damper by Evolving Radial Basis Function Networks. *Eng. Apps. of Art. Intell.*, 19(8):869–881.

Guo, S., Yang, S., and Pan, C. (2006). Dynamical Modeling of Magneto-rheological Damper Behaviors. *Int. Mater, Sys. and Struct.*, 17:3–14.

Li, W. H., Yao, G. Z., and Chen, G. (2000). Testing and Steady State Modeling of a Linear MR Damper under Sinusoidal Loading. *Smart Materials Structures*, 9:95–102.

Lozoya-Santos, J. J., Morales-Menendez, R., and Ramirez-Mendoza, R. (2009). Design of Experiments for MR Damper Modelling. To appear in *Neural Netwotks, Int. Joint Conf. on, IEEE Proc.*

Nino-Juarez, E., Morales-Menendez, R., Ramirez-Mendoza, R., and Dugard, L. (2008). Minimizing the Frequency in a Black Box Model of a MR Damper. In *11th Mini Conf on Vehicle Sys. Dyn., Ident. and Anomalies*.

Savaresi, S. M., Bittanti, S., and Montiglio, M. (2005). Identification of Semi-Physical and Black-Box Non-Linear Models: the Case of MR-Dampers for Vehicles Control. *Automatica,*, 41(1):113–127.

Shivaram, A. C. and Gangadharan, K. V. (2007). Statistical Modeling of a MR Fluid Damper using the Design of Experiments Approach. *Smart Mater. and Struct.*, 16(4):1310–1314.

Spencer, B., Dyke, S., Sain, M., and Carlson, J. (1996). Phenomenological Model of a MR Damper. *ASCE J of Eng Mechanics*.

Wang, D.-H. and Liao, W.-H. (2001). Neural Network Modeling and Controllers for Magneto-Rheological Fluid Dampers. In *Fuzzy Sys.. The 10th IEEE Int. Conf. on*, volume 3, pages 1323–1326.

Wang, D. H. and Liao, W. H. (2005). Modeling and Control of Magnetorheological Fluid Dampers using Neural Networks. *Smart Mater. Struct.*, 14:111–126.

Wang, L. X. and Kamath, H. (2006). Modelling Hysteretic Behaviour in MR Fluids and Dampers using Phase-Transition Theory. *Smart Mater. Struct.*, 15:1725–1733.

SYNCHRONIZATION OF MODIFIED CHUA'S CIRCUITS IN STAR COUPLED NETWORKS

O. R. Acosta del Campo
Engineering Faculty, UABC, Km. 103 Carretera Tij-Ens, Ensenada, México
trecexp13@hotmail.com

C. Cruz-Hernández
Electronics and Telecommunications Department, CICESE, Km. 107 Carretera Tij-Ens, Ensenada, México
ccruz@cicese.mx

R. M. López-Gutiérrez , E. E. García-Guerrero
Engineering Faculty, UABC, Km. 103 Carretera Tij-Ens, Ensenada, México
roslopez@uabc.mx, eegarcia@uabc.mx

Keywords: Hyperchaos synchronization, Modified fourth-order Chua's circuit, Generalized Hamiltonian forms, Observers, Star coupled networks.

Abstract: In this paper, we use Generalized Hamiltonian systems approach to synchronize dynamical networks of modified fourth-order Chua's circuits, which generate hyperchaotic dynamics. Network synchronization is obtained among a single master node and two slave nodes, with the slave nodes being given by observers.

1 INTRODUCTION

The synchronization problem of two chaotic oscillators has received a lot of attention in last decades, see e.g. this example in order to achieve the highest quality possible (Pecora and Carroll, 1990); (Nijmeijer and Mareels, 1997); (López-Mancilla and Cruz-Hernández, 2005); (López-Mancilla and Cruz-Hernández, 2008); (Cruz-Hernández and Nijmeijer, 2000); (Boccaleti and et. al., 2002); (Luo, 2008); (Cruz-Hernández, 2004) and references therein. This interest increases by practical applications in different fields, particularly in secure communications, see e.g. (Cruz-Hernández, 2004); (López-Mancilla and Cruz-Hernández, 2005); (Aguilar-Bustos and Cruz-Hernández, 2008); (Cruz-Hernández and N.Romero-Haros, 2008). Hyperchaotic dynamics characterized by more than one positive Lyapunov exponent are advantageous over simple chaotic dynamics. However, hyperchaos synchronization is a much more difficult problem, see e.g. (Aguilar-Bustos and Cruz-Hernández, 2008) for two coupled oscillators.

In (Posadas-Castillo and et.al.(a), 2007) was developed an experimental study on practical realization to synchronize dynamical networks of Chua's circuits globally coupled. While in recent works (Posadas-Castillo and et. al.(b), 2007); (Posadas-Castillo and et. al., 2008); (H. Serrano Guerrero, 2009) was obtained synchronization in coupled star networks with chaotic nodes given by Nd:YAG lasers and 3*D* CNNs, respectively; by using the approach given in (Wang, 2002). Some literature devoted on synchronization of complex networks (Manrubia and et.al., 2004); (Pogromsky and Nijmeijer, 2001); (Wang, 2002).

Network synchronization of coupled star nodes can be applied to transmit encrypted messages, from a single transmitter to multiple receivers in network communication systems, if the coupled nodes are chaotics. The aim of this paper is to synchronize three modified fourth-order Chua's circuits (which exhibit hyperchaotic behavior) studied in (Thamilmaran and et.al., 2004) in star coupled networks via Generalized Hamiltonian forms and observer design proposed in (Sira-Ramírez and Cruz-Hernández, 2001). This approach presents several advantages over the existing synchronization methods reported in the current literature.

2 PROBLEM SETTING

Consider the following set of *N* interconnected identical dynamical systems

$$x_i = f(x_i) + u_i, \quad i = 1,2,...,N, \tag{1}$$

where $x_i = (x_{i1}, x_{i2}, ..., x_{in})^T \in \mathbb{R}^n$ is the state vector and $u_i = u_{i1} \in \mathbb{R}$ is the input signal of the system i, defined by

$$u_{i1} = c \sum_{j=1}^{N} a_{ij}\Gamma \mathbf{x}_j, \quad i = 1,2,\ldots,N, \tag{2}$$

the constant $c > 0$ represents the *coupling strength*, and $\Gamma \in \mathbb{R}^{n\times n}$ is a constant 0-1 matrix linking coupled states.Whereas, $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{n\times n}$ is the *coupling matrix*, which represents the coupling configuration in (1)-(2). If there is a connection between node i and node j, then $a_{ij} = 1$; otherwise, $a_{ij} = 0$ for $i \neq j$. Note that, if $u_{i1} = 0$, $i = 1,2,...,N$, in (1) we have a set of N isolated dynamical systems, operating with their own dynamics. While, if $u_{i1} \neq 0$ the set constitutes a *dynamical network* and each dynamical system i is called *nodo i*; and under appropiates u_{i1} the dynamical networks can be achieve collective behaviors. It is clear that, the input singal u_{i1} determines the kind of coupling among nodes in the networks. The coupling matrix for star coupled networks is given by

$$A = \begin{pmatrix} N-1 & -1 & -1 & \cdots & -1 \\ -1 & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ -1 & 0 & 0 & \cdots & 0 \\ -1 & 0 & 0 & \cdots & 0 \end{pmatrix} \tag{3}$$

The star coupled configuration for N nodes is shown in Fig. 1, with the common or central node 1.

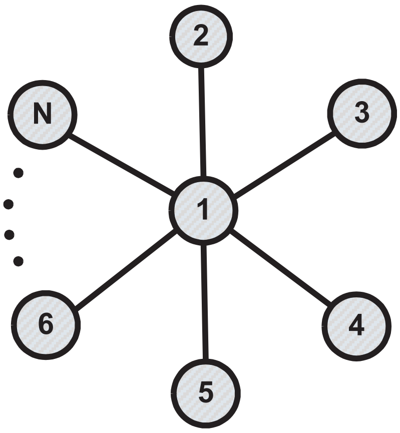


Figure 1: Star coupled configuration with N nodes.

The complex dynamical network (1) is said to achieve (asymptotically) *synchronization*, if (Wang 2002):

$$x_1(t) = x_2(t) = ... = x_N(t) \text{ as } t \to \infty. \tag{4}$$

The synchronization state in (1) can be an equilibrium point, a periodic orbit or, a chaotic attractor. This paper addresses the synchronization problem of dynamical networks (1) with coupled nodes in star topologies. In particular, by choosing a master node with the objective of to impose a particular collective behavior in (1). For illustrative purposes only, we consider three isolated nodes ($N = 3$) to be synchronized (which are described in Section 4), see Fig. 2(a). In Fig. 2(b) is shown this dynamical network with master node $N1$ and two slave nodes $N2$ and $N3$. Our objective is the synchronization of this network, when the coupled nodes are given by modified fourth-order Chua's circuits, to be described in Section 4. This particular coupling topology is important for its application to network communication systems, to transmit messages from a single transmitter to multiple receivers (Chow T.W.W. and Ng, 2001).

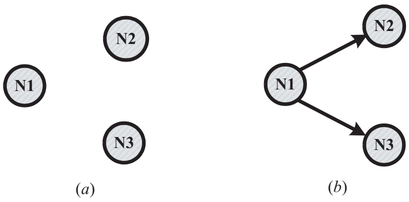


Figure 2: (a) Three isolated nodes. (b) Star coupled network with master node N1.

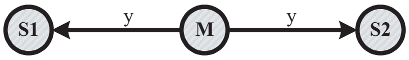


Figure 3: Single master node M and two slave nodes S1 and S2 configuration.

3 SYNCHRONIZATION VIA HAMILTONIAN FORMS

To solve the network synchronization problem stated in previous section, we appeal to synchronization (of two chaotic oscillators) via Hamiltonian forms and observer design reported in (Sira-Ramírez and Cruz-Hernández, 2001). In the sequel, we show that this approach is appropriate to synchronize a coupled star network with three nodes shown in Fig. 2(b). By using the proposed synchronization scheme shown in Fig. 3, where M is given in Hamiltonian form (Eq. (7)) and $S1$ and $S2$ being two observers for M given by Eq. (8).

Consider the following isolated dynamical system

$$\dot{x} = f(x), \tag{5}$$

where $x(t) \in \mathbb{R}^n$ is the state vector, $f : \mathbb{R}^n \to \mathbb{R}^n$ is a nonlinear function.

In (Sira-Ramírez and Cruz-Hernández, 2001) is reported how the dynamical system (5) can be written in the following Generalized Hamiltonian canonical form,

$$\dot{x} = \mathcal{J}(x)\frac{\partial H}{\partial x} + \mathcal{S}(x)\frac{\partial H}{\partial x} + \mathcal{F}(x), \quad x \in \mathbb{R}^n, \tag{6}$$

$H(x)$ denotes a smooth energy function which is globally positive definite in $\mathbb{R}^n$. The gradient vector of H, denoted by $\partial H/\partial x$, is assumed to exist everywhere. We use quadratic energy function $H(x) = (1/2)x^T Mx$ with M being a, constant, symmetric positive definite matrix. In such case, $\partial H/\partial x = Mx$. The matrices, $\mathcal{J}(x)$ and $S(x)$ satisfy, for all $x \in \mathbb{R}^n$, the properties: $\mathcal{J}(x) + \mathcal{J}^T(x) = 0$ and $S(x) = S^T(x)$. The vector field $\mathcal{J}(x)\partial H/\partial x$ exhibits the conservative part of the system and it is also referred to as the workless part, or work-less forces of the system; and $S(x)$ depicting the working or nonconservative part of the system. For certain systems, $S(x)$ is negative definite or negative semidefinite. Thus, the vector field is considered as the dissipative part of the system. If, on the other hand, $S(x)$ is positive definite, positive semidefinite, or indefinite, it clearly represents, respectively, the global, semi-global, and local destabilizing part of the system. In the last case, we can always (although nonuniquely) descompose such an indefinite symmetric matrix into the sum of a symmetric negative semidefinite. matrix $R(x)$ and a symmetric positive semidefinite matrix $N(x)$. Finally, $F(x)$ represents a locally destabilizing vector field.

In the context of observer design, we consider a special class of Generalized Hamiltonian forms (to be considered as the master node M) with linear output map $y(t)$, given by

$$\begin{aligned} \dot{x} &= \mathcal{J}(y)\frac{\partial H}{\partial x} + (\mathcal{I} + \mathcal{S})\frac{\partial H}{\partial x} + \mathcal{F}(y), \; x \in \mathbb{R}^n, \quad (7) \\ y &= \mathcal{C}\frac{\partial H}{\partial x}, \quad y \in \mathbb{R}^m, \end{aligned}$$

where S is a constant symmetric matrix, not necessarily of definite sign. The matrix I is a constant skew symmetric matrix, and $\mathcal{C}$ is a constant matrix.

We denote the estimates of the state $x(t)$ by $\hat{x}_i(t)$, $i = 1,2$ and consider the Hamiltonian energy function $H(\hat{x}_i)$ to be the particularization of H in terms of $\hat{x}_i(t)$. Similarly, we denote by $\eta_i(t)$, $i - 1,2$ the estimated outputs, computed in terms of the estimated states $\hat{x}_i(t)$. The gradient vector $\partial H(\hat{x}_i)/\partial \hat{x}_i$ is naturally, of the form $M\hat{x}_i$ with M being a, constant, symmetric positive definite matrix.

Two *nonlinear state observers* for M (7) are given by

$$\begin{aligned} \dot{\hat{x}}_i &= \mathcal{J}(y)\frac{\partial H}{\partial \hat{x}_i} + (\mathcal{I} + \mathcal{S})\frac{\partial H}{\partial \hat{x}_i} + \mathcal{F}(y) + K_i(y - \eta_i), \quad (8) \\ \eta_i &= \mathcal{C}\frac{\partial H}{\partial \hat{x}_i}, \quad \eta_i \in \mathbb{R}^m, \quad i = 1,2, \end{aligned}$$

with $\hat{x}_i \in \mathbb{R}^n$ and K_i is the observer gain.

The state estimation errors, defined as $e_i(t) = x(t) - \hat{x}_i(t)$ and the output estimation error, defined as $e_{iy}(t) = y(t) - \eta_i(t)$, are governed by

$$\begin{aligned} \dot{e}_i &= \mathcal{J}(y)\frac{\partial H}{\partial e_i} + (\mathcal{I} + \mathcal{S} - K\mathcal{C})\frac{\partial H}{\partial e_i}, \; e_i \in \mathbb{R}^n, \quad (9) \\ e_{iy} &= \mathcal{C}\frac{\partial H}{\partial e_i}, \quad e_{iy} \in \mathbb{R}^m, \quad i = 1,2, \end{aligned}$$

where the vectors $\partial H/\partial e_i$ actually stands, with some abuse of notation, for the gradient vector of the modified energy functions, $\partial H(e_i)/\partial e_i = \partial H/\partial x - \partial H/\partial \hat{x}_i = M(x - \hat{x}_i) = Me_i$. We set, when needed, $I + S = \mathcal{W}$.

A necessary and sufficient condition for global asymptotic stability to zero of the estimation errors (9) is given by the following theorem.

Theorem 1 (Sira-Ramírez and Cruz-Hernández, 2001). The state $x(t)$ of the master node M (7) can be globally, exponentially, asymptotically estimated, by the states $\hat{x}_i(t)$, $i = 1,2$ of the observers (8) if and only if, there exist constant matrices K_i such that the symmetric matrices

$$\begin{aligned} [\mathcal{W} - K_i\mathcal{C}] + [\mathcal{W} - K_i\mathcal{C}]^T &= [\mathcal{S} - K_i\mathcal{C}] + [\mathcal{S} - K_i\mathcal{C}]^T \\ &= 2\left[\mathcal{S} - \frac{1}{2}(K_i\mathcal{C} + \mathcal{C}^T K_i^T)\right] \end{aligned}$$

are negative definite.

4 HYPERCHAOTIC CHUA'S CIRCUIT LIKE NODE

Consider the modified fourth-order Chua's circuit described by (Thamilmaran and et.al., 2004):

$$\begin{aligned} \dot{x}_1 &= \alpha_1(x_3 - f(x_1)), \\ \dot{x}_2 &= -\alpha_2 x_2 - x_3 - x_4, \\ \dot{x}_3 &= \beta_1(x_2 - x_1 - x_3), \\ \dot{x}_4 &= \beta_2 x_2, \end{aligned} \tag{10}$$

with nonlinear function given by

$$f(x_1) = bx_1 + \frac{1}{2}(a - b)(|x_1 + 1| - |x_1 - 1|). \tag{11}$$

With the paramerter values: $\alpha_1 = 2.1429$, $\alpha_2 = -12.83$, $\beta_1 = 0.0393$, $\beta_2 = 0.0015$, $a = -0.0299$, and

$b = 1.995$ the modified Chua's circuit (10)-(11) exhibits hyperchaotic behavior, with two positive Lyapunov exponents. By using the initial conditions $x(0) = (1.1, 0.1, -0.5, 0.01)$, Figs. 1, 2, 3, and 4 show the hyperchaotic attractors $x_1$vs x_2, $x_2$vs x_3, $x_3$vs x_4, and $x_1$vs x_4, respectively.

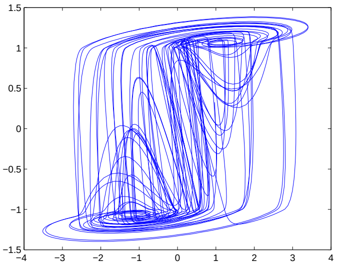


Figure 4: Hyperchaotic attractor projected onto the (x_1, x_2)-plane.

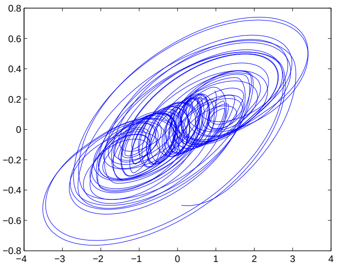


Figure 5: Hyperchaotic attractor projected onto the (x_2, x_3)-plane.

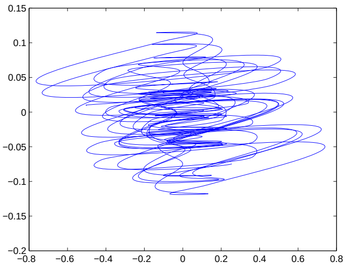


Figure 6: Hyperchaotic attractor projected onto the (x_3, x_4)-plane.

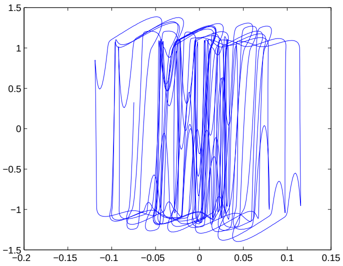


Figure 7: Hyperchaotic attractor projected onto the (x_1, x_4)-plane.

Next, we show the arrangement for star dynamical network by using as coupled node to hyperchaotic Chua's circuit defined by (10)-(11).

5 SYNCHRONIZATION OF HYPERCHAOTIC CHUA'S CIRCUITS IN A STAR NETWORK

In this section, we show the synchronization of three hyperchaotic Chua's circuits in a star coupled network, via Generalized Hamiltonian forms and observer design proposed in (Sira-Ramírez & Cruz-Hernández 2001). Firstly, we rewrite the modified fourth-order Chua's circuit (10)-(11) for the master node as follows.

Taking as Hamiltonian energy function to

$$H(x) = \frac{1}{2}\left(\frac{1}{\alpha_1}x_1^2 + x_2^2 + \frac{1}{\beta_1}x_3^2 + \frac{1}{\beta_2}x_4^2\right). \qquad (12)$$

Modified fourth-order Chua's circuit (10)-(11) in Generalized Hamiltonian form (**master node**, M) according to Eq. (7) is given by

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \end{pmatrix} = \begin{pmatrix} 0 & 0 & \alpha_1\beta_1 & 0 \\ 0 & 0 & -\beta_1 & -\beta_2 \\ -\alpha_1\beta_1 & \beta_1 & 0 & 0 \\ 0 & \beta_2 & 0 & 0 \end{pmatrix} \frac{\partial H}{\partial x} + \qquad (13)$$
$$\begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & -\alpha_2 & 0 & 0 \\ 0 & 0 & -\beta_1^2 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \frac{\partial H}{\partial x} + \begin{pmatrix} -\alpha_1 f(x_1) \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

The destabilizing vector field calls for $x_1(t)$ to be used as the output $y(t)$, of the master node M (13). The matrices C, S, and I are given by

$$C^T = \begin{pmatrix} \alpha_1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad S = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & -\alpha_2 & 0 & 0 \\ 0 & 0 & -\beta_1^2 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$
$$I = \begin{pmatrix} 0 & 0 & \alpha_1\beta_1 & 0 \\ 0 & 0 & -\beta_1 & -\beta_2 \\ -\alpha_1\beta_1 & \beta_1 & 0 & 0 \\ 0 & \beta_2 & 0 & 0 \end{pmatrix}.$$

Next, we design two state observers (slave nodes $S1$ and $S2$, see Fig. 3) for master node (13). The first nonlinear state observer for the Generalized Hamiltonian system (13) (according to Eq. (8) as **slave node**

S1 is given by

$$\begin{pmatrix}\dot{\hat{x}}_{11}\\ \dot{\hat{x}}_{12}\\ \dot{\hat{x}}_{13}\\ \dot{\hat{x}}_{14}\end{pmatrix} = \begin{pmatrix}0&0&\alpha_1\beta_1&0\\ 0&0&-\beta_1&-\beta_2\\ -\alpha_1\beta_1&\beta_1&0&0\\ 0&\beta_2&0&0\end{pmatrix}\frac{\partial H}{\partial \hat{x}} + \begin{pmatrix}0&0&0&0\\ 0&-\alpha_2&0&0\\ 0&0&-\beta_1^2&0\\ 0&0&0&0\end{pmatrix}\frac{\partial H}{\partial \hat{x}} + \begin{pmatrix}-\alpha_1 f(x_1)\\ 0\\ 0\\ 0\end{pmatrix} + \begin{pmatrix}k_{11}\\ k_{12}\\ k_{13}\\ k_{14}\end{pmatrix} e_{1y}, \tag{14}$$

$$\eta_1 = \hat{x}_{11},$$

the second state observer (**slave S2**) is described by

$$\begin{pmatrix}\dot{\hat{x}}_{21}\\ \dot{\hat{x}}_{22}\\ \dot{\hat{x}}_{23}\\ \dot{\hat{x}}_{24}\end{pmatrix} = \begin{pmatrix}0&0&\alpha_1\beta_1&0\\ 0&0&-\beta_1&-\beta_2\\ -\alpha_1\beta_1&\beta_1&0&0\\ 0&\beta_2&0&0\end{pmatrix}\frac{\partial H}{\partial \hat{x}} + \begin{pmatrix}0&0&0&0\\ 0&-\alpha_2&0&0\\ 0&0&-\beta_1^2&0\\ 0&0&0&0\end{pmatrix}\frac{\partial H}{\partial \hat{x}} + \begin{pmatrix}-\alpha_1 f(x_1)\\ 0\\ 0\\ 0\end{pmatrix} + \begin{pmatrix}k_{21}\\ k_{22}\\ k_{23}\\ k_{24}\end{pmatrix} e_{2y}, \tag{15}$$

$$\eta_2 = \hat{x}_{21},$$

where $e_{1y} = x_1 - \hat{x}_{11}$ ($e_{11} = y - \eta_1$) and $e_{2y} = x_1 - \hat{x}_{21}$ ($e_{21} = y - \eta_2$). From master node (13) and slave nodes (14) and (15), we have that the synchronization error dynamics among the master node and two slave nodes (observers) is governed by

$$\begin{pmatrix}\dot{e}_{i1}\\ \dot{e}_{i2}\\ \dot{e}_{i3}\\ \dot{e}_{i4}\end{pmatrix} = \begin{pmatrix}0&\frac{k_{i2}\alpha_1}{2}&\gamma_i&\frac{k_{i4}\alpha_1}{2}\\ -\frac{k_{i2}\alpha_1}{2}&0&-\beta_1&-\beta_2\\ -\gamma_i&\beta_1&0&0\\ -\frac{k_{i4}\alpha_1}{2}&\beta_2&0&0\end{pmatrix}\frac{\partial H}{\partial e_i} + \begin{pmatrix}-k_{i1}\alpha_1&\frac{k_{i2}\alpha_1}{2}&-\frac{k_{i3}\alpha_1}{2}&\frac{k_{i4}\alpha_1}{2}\\ -\frac{k_{i2}\alpha_1}{2}&-\alpha_2&0&0\\ -\frac{k_{i3}\alpha_1}{2}&0&-\beta_1^2&0\\ -\frac{k_{i4}\alpha_1}{2}&0&0&0\end{pmatrix}\frac{\partial H}{\partial e_i} \tag{16}$$

where $\gamma_i = \alpha_1\beta_1 + \frac{k_{i3}\alpha_1}{2}$. Where the synchronization errors are defined by e_1 and e_2 among master M and slaves 1 and 2, respectively. One may now choose the observer gains $K_i = (k_{i1}, k_{i2}, k_{i3}, k_{4i})^T$, $i = 1,2$ in order to guarantee asymptotic exponential stability to zero of the synchronization errors $e_i(t) = (e_{i1}(t), e_{i2}(t), e_{i3}(t), e_{i4}(t))$, $i = 1,2$ as will be shown in the next section.

6 SYNCHRONIZATION CONDITIONS

Now, we examine the stability of the synchronization errors (16) for the network constructed with master (13) and two slaves (14) and (15), with modified Chua's circuits as coupled nodes. Thus, we invoke Theorem 1, which guarantees global asymptotic stability to zero of $e_i(t)$, $i = 1,2$. In particular, for modified Chua's circuit, the matrices $2\left[\mathcal{S} - \frac{1}{2}(K_i\mathcal{C} + \mathcal{C}^T K_i^T)\right]$, $i = 1,2$ shown in Theorem 1, are give by

$$\begin{pmatrix}-2k_{i1}\alpha_1&-k_{i2}\alpha_1&-k_{i3}\alpha_1&-k_{i4}\alpha_1\\ -k_{i2}\alpha_1&-2\alpha_2&0&0\\ -k_{i3}\alpha_1&0&-2\beta_1^2&0\\ -k_{i4}\alpha_1&0&0&0\end{pmatrix}, \; i = 1,2 \tag{17}$$

by applying the Sylvester's Criterion -which provides a test for negative definite of a matrix- thus, we have that the mentioned matrices will be negative definite matrices, if we choose $K_i = (k_{i1}, k_{i2}, k_{i3}, k_{i4})^T$, $i = 1,2$ such that the following conditions are satisfied:

$$\begin{aligned} k_{i1} &\leq 1, \quad (18)\\ 4k_{i1}\alpha_1\alpha_2 - k_{i2}^2\alpha_1^2 &\geq 0,\\ 2\left[\alpha_1\beta_1^2\left(\alpha_1 k_{i2}^2 - 4k_{i1}\alpha_2\right) + k_{i3}^2\alpha_1^2\alpha_2\right] &\geq 0,\\ k_{i4} &= 0. \end{aligned}$$

We have used $K_1 = (3.3, 1.5, 0.39, 0)^T$ and $K_2 = (2.3, 1, 0.3, 0)^T$ with initial conditions: for M, $x(0) = (1.1, 0.1, -0.5, 0.01)$ and for $S1$, $\hat{x}_1(0) = (0.5, 0.3, -0.4, 0)$ and for $S2$, $\hat{x}_2(0) = (1, 0, -0.2, 0.04)$. Fig. 8 shows the synchronization among master node (13) and two slave nodes (14) and (15).

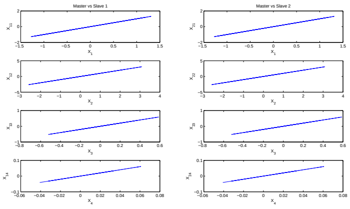


Figure 8: Complete synchronization among states of hyperchaotic master node M and slave nodes S1 and S2.

7 CONCLUSIONS

In this paper, we have presented multiple synchronization of coupled modified fourth-order Chua's circuit, in particular by using star coupled networks.

We have achieve synchronization of three hyperchaotic Chua's circuit (used as fundamental node) in star complex networks, via Generalized Hamiltonian forms and observer design considering a single master node and two slave nodes. This result is particularly interesting given its application in communication network systems, where is required that a single sender transmits simultaneously information to many receivers via a public channel.

ACKNOWLEDGEMENTS

This work was supported by CONACYT, México under Research Grant Nos. J49593-Y, and P50051-Y.

REFERENCES

Aguilar-Bustos, A. and Cruz-Hernández, C. (2008). Synchronization of discrete-time hyperchaotic systems: An application in communications. In *Chaos, Solitons and Fractals*. INSTICC Press.

Boccaleti, S. and et. al. (2002). The synchronization of chaotic systems. In *Physics Reports, 336:1.*

Chow T.W.W., F. J.-C. and Ng, K. (2001). Chaotic network synchronization with applications to communication. In *J. Commun. Syst., 14, 217-30.*

Cruz-Hernández, C. (2004). Synchronization of time-delay chua's oscillator with application to secure communication. In *Nonlinear Dyn. Syst. Theory 4(1), 1-13.*

Cruz-Hernández, C. and Nijmeijer, H. (2000). Synchronization through filtering. In *Int. J. Bifurc. Chaos 10 (4), 763-775.*

Cruz-Hernández, C. and N.Romero-Haros (2008). Communicating via synchronized time-delay chua's circuits. In *Commun. Nonlinear Sci. Numer. Simul.;13(3), 645-59.*

H. Serrano Guerrero, a. e. a. (2009). Synchronization in star coupled networks of 3d cnns and its application in communications. In *submitted to chapter in Evolutionary Design of Intelligent Systems in Modeling', Simulation and Control*. Springer-Verlag.

López-Mancilla, D. and Cruz-Hernández, C. (2005). Output synchronization of chaotic systems: model-matching approach with application to secure communication. In *Nonlinear Dyn. Syst. Theory 5(2), 141-156.*

López-Mancilla, D. and Cruz-Hernández, C. (2008). Output synchronization of chaotic systems under nonvanishing perturbations. In *Chaos, Solitons and Fractals 37, 1172-1186.*

Luo, A. C. J. (2008). A theory for synchronization of dynamical systems. In *Commun. Nonlinear Sci. Numer. Simulat., doi:10.1016/j.cnsns.2008.07.002.*

Manrubia, S. and et.al. (2004). Emergence of dynamical order, synchronization phenomena in complex systems. In *World Scientific, Lecture Notes in Complex Systems Vol. 2, Singapore.*

Nijmeijer, H. and Mareels, I. (1997). An observer looks at synchronization. In *IEEE Trans. Circ. Syst. I 44(10), 882-890.*

Pecora, L. and Carroll, T. (1990). Synchronization in chaotic systems. In *Phys. Rev. Lett. 64, 821-824.*

Pogromsky, Y. A. and Nijmeijer, H. (2001). Cooperative oscillatory behavior of mutually coupled dynamical systems. In *IEEE Trans. Circ. Syst. I 48(2), 152-162.*

Posadas-Castillo, C. and et. al. (2008). Synchronization in a network of chaotic solid-state nd:yag lasers. In *Procs. of the 17th World Congress IFAC, Seoul, Korea, July 6-11, 1565-1570.*

Posadas-Castillo, C. and et. al.(b) (2007). *Lecture Notes in Artificial Intelligence, No. 4529, Synchronization in arrays of chaotic neural networks, in Foundations of Fuzzy Logic and Soft Computing*. Springer-Verlag, Berlin Heidelberg, 1st edition.

Posadas-Castillo, C. and et.al.(a) (2007). Experimental realization of synchronization in complex networks with chua's circuits like nodes. In *Chaos, Solitons and Fractals, In press doi:10.1016/j.chaos.2007.09.076.*

Sira-Ramírez, H. and Cruz-Hernández, C. (2001). Synchronization of chaotic systems: A generalized hamiltonian systems approach. In *Int. J. Bifurc. Chaos 11(5), 1381-1395.*

Thamilmaran, K. and et.al. (2004). Hyperchaos in a modified canonical chua's circuit. In *Int. J. Bifur. Chaos. 14(1), 221243.*

Wang, X. F. (2002). Complex networks: Topology, dynamics and synchronization. In *Int. J. Bifurc. Chaos 12(5), 885-916.*

SIMPLIFICATION OF ATMOSPHERIC MODELS FOR REAL-TIME WIND FORECAST

Qing-Guo Wang, Zhen Ye and Lihong Idris Lim
Department of Electrical and Computer Engineering, National University of Singapore
10 Kent Ridge Crescent, 119260, Singapore
elewqg@nus.edu.sg

Keywords: Real-time wind forecast, Partial differential equations, MM5, Kalman filter.

Abstract: In wind energy industry, it is well known that real-time wind forecast can improve the performance of wind turbines if the prediction information is well used to compensate the uncertainty of the wind. Unfortunately, neither model nor method is available to give a real-time forecast of wind so far. This paper proposed a real-time wind forecast model by simplifying existing weather forecast model, MM5. Details on model simplification, forecast error correction as well as other issues like boundary conditions and simulations are also discussed.

1 INTRODUCTION

Owing to increasing concern over the global environment, there is much interest throughout the world in renewable energy, of which one of the most promising is wind power due to its mature technology, low cost and less environmental impact. Unlike the normal electrical power generation using generated water steam with certain temperature and pressure, wind power utilizes natural but uncertain wind. The wind uncertainty is the root cause for most of the issues in wind power systems, such as nonlinearity, coupling, interaction, and so on. Therefore, it would be much helpful to improve the performance of wind turbines if we could predict the wind and take actions in advance. It would be better if the prediction is real-time since wind is varying all the time.

A natural thought for wind prediction is to make use of weather forecasting models, which has been developed since 1970s and now achieves good prediction for wind, temperature, pressure, moisture, and other weather conditions. Actually in wind power prediction, weather forecasting model has already been applied, see (Landberg, 1999; Joensen et al., 1999; Kazuhito et al., 2006) and references there in, but none of them can give real-time predictions. To the best of our knowledge, not much work has been done yet so far in the real-time wind forecast for wind turbines. This is because:

1. Weather forecasting model is developed for a long-term and large scale forecast, which is not suitable for wind prediction in wind energy industry where only a short-term and small scale forecast is only required;
2. Due to model complexity, the highest temporal resolution of current weather forecasting model is hourly, which is hardly used for real-time prediction.
3. Weather forecasting model lacks of the scheme of correcting prediction error, which is much needed in wind prediction for wind energy industry, especially for real-time forecast.

This paper aims to find a suitable forecasting model for real-time wind prediction. Based on the Fifth-Generation NCAR/Penn State Mesoscale model (MM5) for weather forecast, all possible methods of simplification are discussed to achieve the real-time forecast. Ideas of Kalman filter used to correct the forecast error are also addressed as well as issues on boundary conditions and simulations.

2 MM5 FORECASTING MODEL

MM5 forecasting model is the latest in a series developed from a mesoscale model used by Anthes at Penn State in the early 1970s that was later documented by (Anthes and Warner, 1978). Since that time, it has undergone many changes designed to broaden its applications, including (i) a multiple-nest capability; (ii) nonhydrostatic dynamics; (iii) a four-dimensional data assimilation (Newtonian nudging) capability; (iv) increased number of physics options; and (v) portability to a wider range of computer platforms.

In terms of terrain following coordinates (x,y,σ), the partial differential equations for the nonhydrostatic model's basic variables excluding moisture are:

Pressure

$$\frac{\partial p'}{\partial t} - \rho_0 g w + \gamma p \nabla \cdot \mathbf{V} = \mathbf{V} \cdot \nabla p' + \frac{\gamma p}{T}\left(\frac{\dot{Q}}{c_p} + \frac{T_0}{\theta_0} D_\theta\right), \tag{1}$$

Momentum (x-component)

$$\frac{\partial u}{\partial t} + \frac{m}{\rho}\left(\frac{\partial p'}{\partial x} - \frac{\sigma}{p^*}\frac{\partial p^*}{\partial x}\frac{\partial p'}{\partial \sigma}\right) = -\mathbf{V}\cdot\nabla u$$
$$+ v\left(f + u\frac{\partial m}{\partial y} - v\frac{\partial m}{\partial x}\right) - ew\cos\alpha - \frac{uw}{r} + D_u, \tag{2}$$

Momentum (y-component)

$$\frac{\partial v}{\partial t} + \frac{m}{\rho}\left(\frac{\partial p'}{\partial y} - \frac{\sigma}{p^*}\frac{\partial p^*}{\partial y}\frac{\partial p'}{\partial \sigma}\right) = -\mathbf{V}\cdot\nabla v$$
$$- u\left(f + u\frac{\partial m}{\partial y} - v\frac{\partial m}{\partial x}\right) + ew\sin\alpha - \frac{vw}{r} + D_v, \tag{3}$$

Momentum (z-component)

$$\frac{\partial w}{\partial t} + \frac{\rho_0}{\rho}\frac{g}{p^*}\frac{\partial p'}{\partial \sigma} + \frac{g}{\gamma}\frac{p'}{p} = -\mathbf{V}\cdot\nabla w + g\frac{p_0}{p}\frac{T'}{T_0}$$
$$- \frac{gR_d}{c_p}\frac{p'}{p} + e(u\cos\alpha - v\sin\alpha) + \frac{u^2+v^2}{r} + D_w, \tag{4}$$

Thermodynamics

$$\frac{\partial T}{\partial t} = -\mathbf{V}\cdot\nabla T + \frac{1}{\rho c_p}\left(\frac{\partial p'}{\partial t} + \mathbf{V}\cdot\nabla p' - \rho_0 g w\right)$$
$$+ \frac{\dot{Q}}{c_p} + \frac{T_0}{\theta_0} D_\theta, \tag{5}$$

where p, ρ, T are pressure (Pa), density $(\mathrm{kg}\cdot\mathrm{m}^{-3})$, and temperature (K), respectively. The subscript "0" represents the reference-state. u, v, w are component of wind velocity $(\mathrm{m}\cdot\mathrm{s}^{-1})$ in eastward, northward, and vertical direction, respectively. Q is diabatic heating rate per unit mass $(\mathrm{J}\cdot\mathrm{kg}^{-1}\cdot\mathrm{s}^{-1})$. c_p is specific heat at constant pressure for dry air. $\gamma = c_p/(c_p - R)$ is ratio of heat capacities. $R = 287\mathrm{J}\cdot\mathrm{kg}^{-1}\cdot\mathrm{K}^{-1}$ is ideal gas constant. θ is potential temperature (K). D_A is diffusion and PBL tendency for variable A. m is map-scale factor. $p' = p - p_0$ is perturbation pressure (Pa). $p^* = p_s - p_t$, p_s and p_t are surface and top pressures respectively of the reference state. $\sigma = (p_0 - p_t)/(p_s - p_t)$ is nondimensional vertical coordinate of model. f is Coriolis parameter, $e = 2\Omega\cos\lambda$, $\alpha = \phi - \phi_c$, Ω is angular velocity of the earth, λ is latitude, ϕ is longitude, and ϕ_c is central longitude. r is the radius of the earth.

$$\mathbf{V}\cdot\nabla A \equiv mu\frac{\partial A}{\partial x} + mv\frac{\partial A}{\partial y} + \dot{\sigma}\frac{\partial A}{\partial \sigma}, \tag{6}$$

$$\dot{\sigma} = -\frac{\rho_0 g}{p^*}w - \frac{m\sigma}{p^*}\frac{\partial p^*}{\partial x}u - \frac{m\sigma}{p^*}\frac{\partial p^*}{\partial y}v, \tag{7}$$

$$\nabla\cdot\mathbf{V} = m^2\frac{\partial}{\partial x}\left(\frac{u}{m}\right) - \frac{m\sigma}{p^*}\frac{\partial p^*}{\partial x}\frac{\partial u}{\partial \sigma} + m^2\frac{\partial}{\partial y}\left(\frac{v}{m}\right)$$
$$- \frac{m\sigma}{p^*}\frac{\partial p^*}{\partial y}\frac{\partial v}{\partial \sigma} - \frac{\rho_0 g}{p^*}\frac{\partial w}{\partial \sigma}. \tag{8}$$

The derivations of above model equations (1)-(5) are based on the gas law and first law of thermodynamics. More details can be found in (Grell et al., 1995; Dudhia et al., 2005). Obviously, MM5 is a 5-dimensional model of partial differential equations with structure of coupled variables, which causes its solving time consuming. Currently, the forecast of MM5 can only achieve hourly updation. To implement real-time forecast of wind, simplifications have to be made.

3 SIMPLIFICATION TO MM5 FOR WIND FORECAST

Comparing with weather forecast, wind forecast for wind turbines has its own uniqueness.

1. The rotor blade length of wind turbine is usually less than 150m, so the pressure change along the vertical direction for wind turbine is not too much.
2. The rotation of wind turbine is not driven by the vertical pressure on the blade, but by the horizontal velocity difference on its top and bottom surfaces. Pressure or vertical velocity has less contribution.
3. Temperature may not be a necessary option in real-time forecast as the rotation of wind turbine is not sensitive to temperature change.

Thus, the MM5 model of equations (1)-(5) can be simplified in the following way:

For real-time forecast, the time interval of two continuous predictions should be very small, say one minute. In such a short period, temperature changes can be neglected. Therefore, $\partial T/\partial t = 0$, $\mathbf{V}\cdot\nabla T = 0$ and (5) becomes

$$\frac{\dot{Q}}{c_p} + \frac{T_0}{\theta_0}D_\theta = -\frac{1}{\rho c_p}\left(\frac{\partial p'}{\partial t} + \mathbf{V}\cdot\nabla p' - \rho_0 g w\right). \tag{9}$$

Substituting (9) into (1) yields

$$\frac{\partial p'}{\partial t} - \rho_0 g w + p\nabla\cdot\mathbf{V} = \mathbf{V}\cdot\nabla p', \tag{10}$$

since $p/(\rho T c_p) = 1-\gamma^{-1}$ according to gas law. Thus, (1) is simplified to be (10) and model dimension is reduced as (5) is missing.

As pressure changes along the vertical direction of wind turbine is not too much, $\partial p'/\partial z \approx 0$. By the coordinate transformation $(x,y,z) \to (x,y,\sigma)$,

$$\left(\frac{\partial}{\partial x}\right)_z \to \left(\frac{\partial}{\partial x}\right)_\sigma - \left(\frac{\partial z}{\partial x}\right)_\sigma \frac{\partial}{\partial z}, \tag{11}$$

where $\mathrm{d}z = -\mathrm{d}p_0/(\rho_0 g) = -(p^*\mathrm{d}\sigma + \sigma\mathrm{d}p^*)/(\rho_0 g)$, so

$$\left(\frac{\partial}{\partial x}\right)_z \to \left(\frac{\partial}{\partial x}\right)_\sigma - \frac{\sigma}{p^*}\frac{\partial p^*}{\partial x}\frac{\partial}{\partial \sigma}. \tag{12}$$

Thus,

$$\left(\frac{\partial p'}{\partial x}\right)_z \to \left(\frac{\partial p'}{\partial x}\right)_\sigma - \frac{\sigma}{p^*}\frac{\partial p^*}{\partial x}\frac{\partial p'}{\partial \sigma} = 0, \tag{13}$$

$$\left(\frac{\partial p'}{\partial y}\right)_z \to \left(\frac{\partial p'}{\partial y}\right)_\sigma - \frac{\sigma}{p^*}\frac{\partial p^*}{\partial y}\frac{\partial p'}{\partial \sigma} = 0, \tag{14}$$

and (2) and (3) can be simplified as

$$\frac{\partial u}{\partial t} = -\mathbf{V}\cdot\nabla u + v\left(f + u\frac{\partial m}{\partial y} - v\frac{\partial m}{\partial x}\right) \\ -ew\cos\alpha - \frac{uw}{r} + D_u, \tag{15}$$

$$\frac{\partial v}{\partial t} = -\mathbf{V}\cdot\nabla v - u\left(f + u\frac{\partial m}{\partial y} - v\frac{\partial m}{\partial x}\right) \\ +ew\sin\alpha - \frac{vw}{r} + D_v. \tag{16}$$

If ignoring the effect of vertical velocity w on horizontal momentum, (15) and (16) can be further simplified as

$$\frac{\partial u}{\partial t} = -\mathbf{V}\cdot\nabla u + v\left(f + u\frac{\partial m}{\partial y} - v\frac{\partial m}{\partial x}\right) + D_u, \tag{17}$$

$$\frac{\partial v}{\partial t} = -\mathbf{V}\cdot\nabla v - u\left(f + u\frac{\partial m}{\partial y} - v\frac{\partial m}{\partial x}\right) + D_v. \tag{18}$$

By the above approximation, MM5 model is simplified as (10), (17) and 18) with less dimensions and variables, which is suitable for the real-time prediction.

4 BOUNDARY CONDITIONS

Both MM5 and simplified model are composed of partial differential equations where only numerical solutions are available. Thus, boundary conditions have to be set in addition to initial values prior to running a simulation. Here in our real-time wind forecast, wind velocity, temperature and pressure are specified as boundaries.

The boundary values can come from real-time observations of the wind. In this case, some weather stations have to be set up at the outer place of the wind power plant for measurement. The distance from one station to one wind turbine is a trade-off between prediction accuracy and computation burden. The nearer the station is to the wind turbine, the more accurate the wind prediction, but the less time for solving the equations. Alternatively, the boundary values can come from another model's forecast (in real-time forecasts).

The shape of the boundary can be determined freely, depending on the convenience of computation. Rectangle and circle are two types broadly used. The area surrounded by the boundary is then grided evenly and every grid point gives a wind prediction after solving the model numerically. Normally, the wind turbine should be one grid point inside the boundary. But due to its unevenly distribution, this is usually not the case. Therefore, linear interpolation has to be applied to give the final wind prediction.

5 ERROR CORRECTION

To improve the forecast accuracy, some "feedback" strategy should be introduced for error correction by comparing the estimated and true values. This can be done by Kalman filter. Thus, the loss of accuracy caused by the model simplification can be made up, but computation increases inevitably. Therefore, trade-off should be made between model simplification and Kalman filter design.

6 SIMULATIONS

To verify whether the proposed model can give real-time forecast of wind, a simulation has to be conducted. Two ways are available to this end. One is to use the code of MM5 model, which is available at the web site of National Center for Atmospheric Research (NCAR) for free download. The other one is to apply the Matlab Toolbox of partial differential equation to the proposed model. This part of work is still under study and will supplemented in the final version if possible.

7 CONCLUSIONS

This paper proposed a real-time wind forecast model for wind energy industry by simplifying the existing MM5 model of weather forecast. Details of the simplification method is given as well as other issues on model implementation. To our best knowledge, no similar model is available for wind forecast in wind energy industry so far.

REFERENCES

Anthes, R. A. and Warner, T. T. (1978). Development of hydrodynamic models suitable for air polpollution and other mesometeorological studies. *Mon. Wea. Rev.*, 106:1045–1078.

Dudhia, J., Gill, D., Manning, K., Wang, W., and Bruyere, C. (2005). *PSU/NCAR Mesoscale Modeling System Tutorial Class Notes and Users' Guide (MM5 Modeling System Version 3)*. National Center for Atmospheric Research, USA.

Grell, G. A., Dudhia, J., and Stauffer, D. R. (1995). *A Description of the Fifth-Generation Penn State/NCAR Mesoscale Model (MM5)*. National Center For Atmospheric Research, USA.

Joensen, A., Giebel, G., Landberg, L., Madsen, H., and Nielsen, H. A. (1999). Model output statistics applied to wind power prediction. In *Proceedings of European Wind Energy Conference*, pages 1177–1180, Nice, France.

Kazuhito, F., Jun, Y., Akira, T., Tomonao, K., and Takashi, Y. (2006). Wind power generation forecast using the real-time local weather forecasting system. *Wind Energy*, 30(1):92–98.

Landberg, L. (1999). Short-term prediction of the power production from wind farms. *J. Wind Eng. Ind. Aerodyn.*, 8:207–220.

SAFE CONTROLLERS DESIGN FOR HIBRID PLANTS
The Emergency Stop

Eurico Seabra and José Machado
Mechanical Engineering Department / CT2M, Enginnering School, University of Minho, 4800-058 Guimarães, Portugal
eseabra@dem.uminho.pt, jmachado@dem.uminho.pt

Keywords: Safe Controllers, Emergency Stop, GEMMA, Hybrid Plants.

Abstract: This paper presents and discusses a case study that applies a global approach for considering all the automation systems emergency stop requirements. The definition of all the functioning modes and all the stop tasks of the automation system is also presented according the standards EN 418 and EN 60204-1. All the aspects related with the emergency stop are focused in a particular way. The proposed approach defines and guarantees the safety aspects of an automation system controller related with the emergency stop. For the controller structure it is used the GEMMA formalism; for the controller entire specification it is used the SFC and for the controller behavior simulation it is used the Automation studio software.

1 INTRODUCTION

This work is inserted in a bigger project being developed at the School of Engineering of University of Minho (Portugal) - involving four Departments of the School: the Mechanical Engineering Department, the Electronics Department, the Informatics Department and the Industrial Engineering Department - related with application of several techniques in order to obtain safe controllers for Automation Systems.

The same team of this project has developed another project, before this one, where it were studied aspects relied to plant modeling of timed systems and its influence on the Simulation and Formal Verification of Automation Systems Controllers (Machado *et al*, 2008), (Seabra *et al*, 2007), (Machado and Seabra, 2008).

In the actual study it is intended to study and develop some techniques in order to obtain safe controllers for hybrid plants. The first results are presented on this paper where it is presented the aspects relied with the emergency stop of automation systems and all the aspects to considerer when there are defined the functioning modes and the stop tasks of an automation system (EN 418). Also, the controller, in general, will need to comply with Safety of machines requirements (EN 60204-1).

For the Safety controllers design, there are applied some techniques like synthesis techniques (Ramadge and Wonham, 1987) or analysis techniques (Frey and Litz, 2000) in order to be accomplished the desired specifications for the automation system behavior. Between these techniques there are considered, in more detail, in this paper the analysis techniques.

Considering some aspects and techniques inside of the analysis techniques group the most important are: Identification (Klein, 2004), Simulation (Baresi, 2002) and Formal Verification (Rossi, 2004). This approach is based on Simulation Techniques and it is considered, on the first hand, a discrete controller and the hybrid plant are modeled as being discrete. This simplification will allow us to obtain, faster and with the same rigor, some results relied with the emergency stop behavior for the automation system.

The Emergency Stop is one of the most important aspects attending to the safety of people, goods and equipments that interact with the automation system.

In order to obtain safe controllers, it must obey at some rules (EN 418, 1992), (EN 60204-1, 1997):

- a fault in the software of the control system does not lead to hazardous situations;
- reasonably foreseeable human error during operation does not lead to hazardous situations;
- the machinery must not start unexpectedly;
- the parameters of the machinery must not change in an uncontrolled way, where such change may lead to hazardous situations;

- the machinery must not be prevented from stopping if the stop command has already been given;
- no moving part of the machinery or piece held by the machinery must fall or be ejected;
- automatic or manual stopping of the moving parts, whatever they may be, must be unimpeded;
- the safety-related parts of the control system must apply in a coherent way to the whole of an assembly of machinery and/or partly completed machinery;

As guarantee that the developed controller will react always according the expected behavior, it is only necessary to model the controller and the plant as being discrete. Indeed, our system has a hybrid plant, but the properties of behavior that we intend to guarantee, for our system, are only related with discrete behavior.

For more complex properties – dealing with hybrid behavior of the automation system – it will be necessary to model the controller and the plant as hybrid. This will be done on a next step in this complex research project, using formalisms and tools well adapted for these tasks, like, for instance, Stategraphs (Otter et *al*, 2005) to model the controller and Modelica programming language (Elmqvist and Mattson, 1997) to model the plant.

On this study, presented on this paper, we use the GEMMA (ADEPA, 1992) for the controller structure, the SFC (IEC 60848, 1998) as controller specification formalism and the Automation Studio software (Automation Studio, 2004) for the simulation tasks of the controller specification. With this set of formalisms and tools we demonstrate that it is all we need for guarantee all the desired behavior for the automation system when the emergency stop command is actuated.

In this first approach it is intended to conclude about the more important behavior properties related with the emergency stop of the automation system and the use of the formalisms, and tools, previously described (GEMMA, SFC and Automation Studio) allow us to obtain the desired results in a fast and expedite way.

One of the limitations of this first approach is that the hybrid plant is model as discrete, but this simplification allows the fast obtaining of results related with discrete desired behaviors, being the efforts of modeling more simple and fast.

As we presented before, this step on a more complex approach is only the first step considered in order to guarantee the desired behavior in case of occurrence of the "Emergency" command.

To accomplish the proposed goals, in this work, the paper is organized as follows. In Section 1, it is presented the challenge proposed to achieve in this work. Section 2 presents the case study plant related with an automatic system for filling and encapsulating bottles. Further, it is presented the base controller specification and the total controller structure that includes the emergency stop. Section 3 is exclusively devoted to the emergency stop techniques discussion. Section 4 presents and discusses the emergency stop adopted solution and the total controller specification. Finally, in Section 5, the main conclusions and some future directions to follow in this project that is now starting at the School of Engineering of University of Minho.

2 SYSTEM DESCRIPTION

The case study corresponds to an automatic machine of filling and encapsulating bottles (Fig. 1). This is divided in three modules, transport and feeding, filling and encapsulating. To increase the productivity, is used a conveyor with several alveoli for the bottles to allow the operation in simultaneous of the three modules.

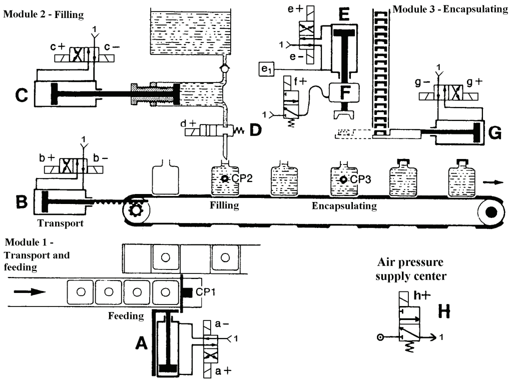


Figure 1: Case study plant.

The transport and feeding module is constituted by a pneumatic cylinder (A) that is the responsible for the bottles feeding of the conveyor and another pneumatic cylinder (B) that executes the step/incremental advance of the conveyor.

The filling module is composed by a volumetric dispenser, a pneumatic cylinder (C) that actuate the dispenser and an on/off valve (D) to open and close the liquid supply.

The encapsulating module has a pneumatic cylinder (G) to feed the cover, a pneumatic motor (F) to screw the cover and a pneumatic cylinder (E) to advance the cover. The cylinder (E) moves forward until the existent cover, it retreats with this cover during the retreat of (G), continuously it moves forward again with rotation of the motor F to screw the cover.

2.1 Base Controller Behaviour Specification

Figure 2 shows the base SFC of the system controller, corresponding only to the "Normal production" mode. The basic sensors involved are: two end-course-sensors for each cylinder (example: cylinder A, sensor a0 and a1, respectively, retreated and advanced) and a sensor of pressure e1, which detects the point of contact/stop of the cylinder E in any point of its course.

The valve D and the motor F don't have position sensor because they are difficult to implement.

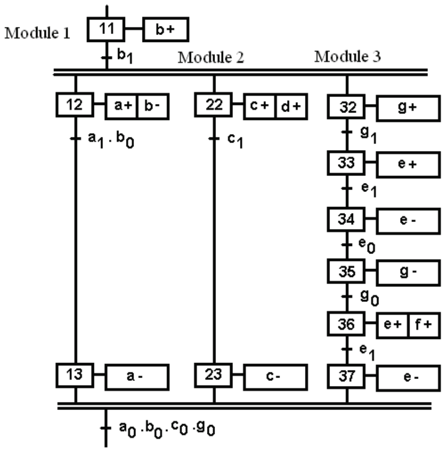


Figure 2: Base SFC specification controller.

On the other hand, in order to obtain the total SFC controller, which includes all the operation modes required for the correct operation of the system, was used the graphic chart of GEMMA because it allows the definition of the run and stop machine tasks.

2.2 Total Controller Behaviour Structure

Figure 3 shows the GEMMA graphic chart developed for the case study presented. The considered tasks are described to proceed:

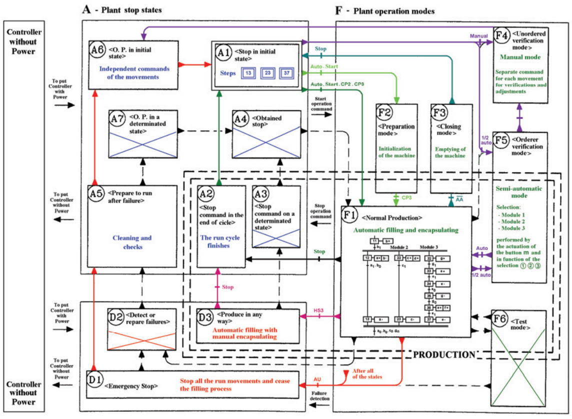

A - Plant stop states
F - Plant operation modes
Controller without Power
To put Controller with Power
To put Controller without Power
A6 <O. P. in initial state>
Independent commands of the movements
A1 <Stop in initial state>
Steps 13 23 37
A7 <O. P. in a determinated state>
A4 <Obtained stop>
A5 <Prepare to run after failure>
Cleaning and checks
A2 <Stop command in the end of cicle>
The run cycle finishes
A3 <Stop command on a determinated state>
Stop
D2 <Detect or repare failures>
D3 <Produce in any way>
Automatic filling with manual encapsulating
D1 <Emergency Stop>
Stop all the run movements and cease the filling process
Manual
Stop
Auto . Start
Auto . Start . CP2 . CP3
Start operation command
Stop operation command
F2 <Preparation mode>
Initialization of the machine
F3 <Closing mode>
Emptying of the machine
F4 <Unordered verification mode>
Manual mode
Separate command for each movement for verifications and adjustments
F5 <Orderer verification mode>
1/2 auto
Semi-automatic mode
Selection:
- Module 1
- Module 2
- Module 3
performed by the actuation of the button m and in function of the selection ① ② ③
CP3
$\overline{AA}$
F1 <Normal Production>
Automatic filling and encapsulating
Auto
1/2 auto
HS3
F6 <Test mode>
PRODUCTION
AU
After all of the states
Failure detection

Figure 3: GEMMA of the plant controller.

A1 – The task A1 "Stop in the initial state" represents the task of the machine represented in the Figure 1.

F1 – Coming of the task A1, when it occurs the start command of the machine, it happens the change for the task F1 "Normal production" (Filling and automatic encapsulating) with the consequent execution of base SFC presented in the figure 2.

A2 – When it happens the stop command of the machine the run cycle finishes in agreement with the condition described at the task A2 “Stop command in the end of cycle”.

F2 – When the machine is "empty" (without bottles) it is necessary to feed bottles progressively, being the machine ready to begin the normal production (task F1) when it has bottles in the conveyor positions of the production modules 2 and 3, respectively. This operation is defined by the task F2 “Preparation mode”.

F3 – The "Closing mode" of the task F3 allows the reverse operation, that is, the progressive stop of the machine with the exit of all of the bottles (emptying of the machine).

D3 – When the encapsulating module is out of service it can be decided to produce in any way, that is, to perform the bottle filling in an automatic way and posterior manual encapsulating, this is main purpose of the task D3 "Production in any way".

D1 – In the case of a situation emergency to occur, the task D1 “Emergency stop" is executed. This stops all the run actions and closes the filling valve to stop the liquid supply.

A5 – After the emergency stop (task D1), the cleaning and the verification are necessary: this is the purpose of the task A5 "Prepare to run after failure“.

A6 – After the procedures of cleaning and verification they be finished becomes necessary to perform the return to the initial task of the machine, as described at the task A6 "O.P. (operative plant) in the initial state".

F4 – For example, to the volume regulation of the bottle liquid dispenser and adjustment of the bottles feeder, a separate command for each movement is required, according to the task F4 "Unordered verification mode".

F5 – For detailed operation checks, a semiautomatic command (only one cycle) it is necessary to check the functioning of each module: task F5 "Ordered verification mode".

To be possible the GEMMA evolution becomes necessary existing transition conditions for the run and stop operation modes, described previously.

These transition conditions will be accomplished using GEMMA, as presented to proceed:

- To allow the progressive feeding demanded in the preparation way (F2) and the progressive discharge required in the closing way (F3) it will be necessary to consider sensors that detect the bottles presence under each one of the modules 1, 2, 3, respectively, CP1, CP2, CP3 (Fig. 1);

- Also, it will be necessary a command panel that supplies the transition conditions given by an operator (Fig. 4).

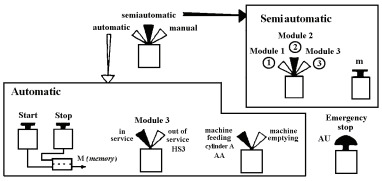


Figure 4: Command panel of the system controller.

In the command panel, there is a main switch that allows selecting the “automatic”, “semiautomatic” and “manual” operations modes.

To the "automatic" option correspond:

- Two buttons "start" and "stop" whose action is memorized in memory M;

- A switch HS3 to put "in service" or "out of service" the module 3;

- A switch AA to control the bottles feeding permission (cylinder A), to allow the emptying of the machine.

These switches/buttons, and sensors CP1, CP2 and CP3, are the transition conditions of the tasks A1, F1, F2, F3, A2 and D3, as shown in Figure 3.

The "semiautomatic" option corresponds to the task F5 "Ordered verification mode", that allows with the actuation of button (m), to check one cycle operation of each modules, selected by the "semiautomatic" switch①,②, or ③.

The "manual" option corresponds to the tasks F4, A5 and A6, which required a separate command from each movement using a direct command on the directional valves.

Finally, the AU button (Emergency stop) allows pass to task D1 starting from all of the tasks.

The implementation of total controller's specification, based on GEMMA presented in figure 3, it can be realized using the following two alternative methods:

- Multiple SFC – develop one SFC for each task;
- Single SFC – develop one SFC for all tasks.

The multiple SFC methodology is represented in figure 5, it includes a high level SFC that translates the GEMMA (main routine) and multiple SFC that correspond to each task (subroutines).

On the other hand, the single SFC method corresponds to the implementation of all GEMMA tasks behaviour in a total SFC. This was the method used in the presented case study (see section 4).

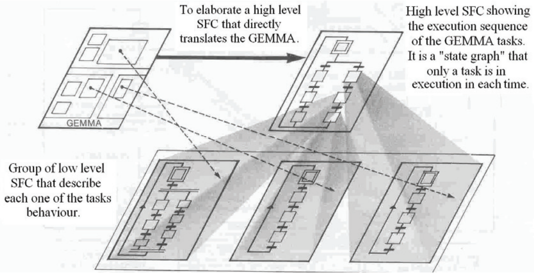

To elaborate a high level SFC that directly translates the GEMMA.
High level SFC showing the execution sequence of the GEMMA tasks. It is a "state graph" that only a task is in execution in each time.
GEMMA
Group of low level SFC that describe each one of the tasks behaviour.

Figure 5: GEMMA implementation with multiple SFC.

3 EMERGENCY STOP

The emergency stop must always change the controller task and it should be obligatorily available in any state of the SFC controller.

The types of emergency stops are divided in two main groups:

- Without emergency sequence - the actuation of the emergency button stops the system/automatism through the inhibition of the outputs and/or for stop the evolution of SFC.

- With emergency sequence - the actuation of the emergency button starts a particular predefined procedure.

3.1 Without Emergency Sequence

The emergency without emergency sequence can be performed in three alternative modes:

- Outputs inhibition;
- Evolution stop,

- Outputs inhibition and evolution stop.

In the case of outputs inhibition the actuation of emergency button doesn't stop by itself the evolution of the SFC controller, but it inhibits the outputs associated to their stages, as shown in the figure 6. The outputs eventually ON (state 1) they are turn OFF (state 0), as well as, usually the evolution of SFC is stopped by the non fulfilment of the receptivity's.

This can be obtained through the insert of inhibition functions in the interface with the machine plant. In this case, after the occurrence of an emergency stop, the actuators command should be particularly well studied in agreement with the type of expected response.

For instance, for the cylinders directional valves:

- One stable state valve (single control with spring return), if it be demanded a cylinder return for a given position.

- Two stable state valve (double control), if it be demanded a stop at the end of the cylinder movement.

- Valve with three positions (double control and spring return), if it be demanded a cylinder stop in the actual position.

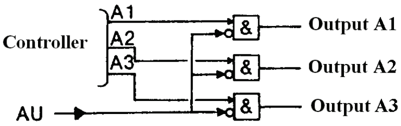


Figure 6: Functional diagram of outputs inhibition.

In the other hand, in the case of evolution stop the condition AU is present in all of the SFC receptivity's (Fig.7a). With the actuation of emergency button AU, no receptivity can be validated and, this way, the controller SFC cannot steps forward. With the AU shutdown a new cycle evolution is allowed.

It is of highlighted that in this situation, the outputs associated to the active stages stay validated. This way, the start movements can continue, which be able to result in dangerous situations and/or to get to a situation that originates a future blockade of the SFC evolution.

Finally, also it is possible to use in simultaneous the two described types of emergency stop without emergency sequence, outputs inhibition and evolution stop (Fig. 7b). This situation is the more used in practice, when if it doesn't turn necessary the use of an emergency sequence. Seen that has the advantage of allowing, after the emergency button shutdown, the pursuit of the evolution of the system starting from the same instant in that it was stopped.

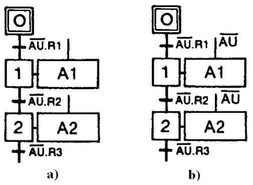


Figure 7: a - Evolution stop; b - Evolution stop and outputs inhibition.

3.2 With Emergency Sequence

This type of emergency implies the introduction of an emergency sequence. Through the activation of the emergency button AU an emergency sequence can be added to the normal run SFC (Fig. 8).

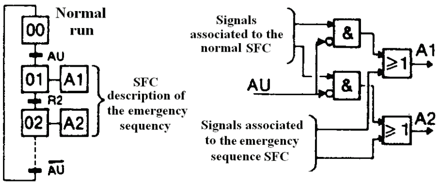

Normal run
00
AU
01
A1
R2
02
A2
$\overline{AU}$
SFC description of the emergency sequency
Signals associated to the normal SFC
AU
Signals associated to the emergency sequence SFC
&
&
≥1
≥1
A1
A2

Figure 8: Introduction of an emergency sequence.

4 EMERGENCY STOP ADOPTED SOLUTION

The emergency stop adopted for the case study presented was obtained according the standards EN 418 and EN 60204-1.

According to the behaviour of the case study was selected the emergency stop with emergency sequence. The considered requirements that should be accomplished by the emergency sequence are:

- Stop all of the movements;

- Stop the filling operation.

To obtain these procedures it was crucial the selection of the type of the directional valves appropriate to accomplish in simultaneous the requirements of the emergency stop and the plant behaviour.

The directional valves specifications used were the type of control (single solenoid control with spring return or double solenoid control) and number of ways/ports.

The first security requirement referred, related with the stop of the movements, was obtained by stopping the air compressed supply to the directional valves of the cylinders A, B, C, E, G and of the motor F. For that, as shown in figure 1, the air supply will be centralized and controlled through a directional valve 3/2 way normally closed with spring return (H).

The second security requirement, related with the stop the filling operation, was performed through the turn OFF of the filling directional valve 2/2 way normally closed with spring return (D).

The figure 9 shows the total controller SFC specification based on the GEMMA implementation with the single SFC method (see section 2.2).

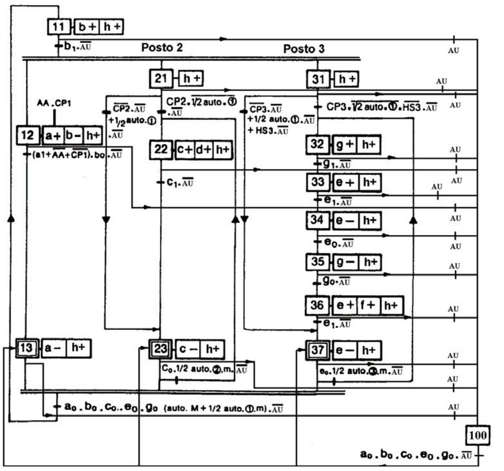


Figure 9: Total SFC controller specification with emergence sequence.

All the controller specification, presented on the previous figure, was simulated on Automation Studio Software. The obtained results leaded to the conclusions that all the requirements defined on the Emergency Stop Standards, were accomplished.

Further, the specification was translated to Ladder Diagrams according to the SFC algebraic formalization and implemented on a Programmable Logic Controller (PLC) adopted as the controller physical device. This part of the developed work is not detailed on this publication.

5 CONCLUSIONS

It was presented, on a systematic way, the adopted techniques for the emergency stop behavior specification of automation systems.

The ways to translate the GEMMA graphical chart to the low level specification was also presented and discussed.

The standards (EN418, EN60204-1) related with the stop emergency specifications were considered and all the requirements were accomplished.

The obtained results, by simulation with Automation Studio software, show that the adopted approach is adequate.

Further work will be devoted, in one hand, to the application of formal methods to verify some important system's behavior (taking into account the discrete behavior of the system) and, in other hand, the application of modeling techniques for hybrid systems and respective tools for simulation and formal verification.

REFERENCES

ADEPA, 1992. *GEMMA (2nd Edition) – Guide d`Étude des Modes de Marches et d`Arrêts*.

Automation Studio, 2004. *Famic Technologies inc., Automation Studio 5.0*, http://www.automationstudio.com.

Baresi L., Mauri M., Pezzè M., 2002. PLCTools: Graph Transformation Meets PLC Design. *Electronic Notes in Theoretical Computer Science*, Vol. 72, No. 2.

Elmqvist E., Mattson S., 1997. *An Introduction to the Physical Modelling Language Modelica*. ESS'97. Passau, Germany.

EN 418, 1992. Safety of machinery. Emergency stop equipment, functional aspects. Principles for design. *European Standard*.

EN 60204-1, 1997. Safety of Machinery - Electrical Equipment of Machines - Part 1: General Requirements-IEC 60204-1. *European Standard*.

Frey G., Litz L., 2000. *Formal methods in PLC programming*. IEEE Conference on Systems, Man and Cybernetics, SMC 2000, Nashville, October 8-11.

IEC 60848, 1998. *Specification language GRAFCET for sequential function chart*. ed. 2.

Klein S., 2005. *Fault detection of discrete event systems using an identification approach*. PhD Thesis, University of Kaiserslautern.

Machado, J., Seabra, E. A. R., 2008. *Real-Time Systems Safety Control considering Human-Machine Interface*. ICINCO'2008, May 10-14, Funchal, Madeira, Portugal,. 6p.

Machado, J., Seabra, E. A. R., Campos, J., Soares, F. O., Leão, C. P., Silva, J. C. L., 2008. Simulation and Formal Verification of Industrial Systems Controllers. *Symp. Series in Mech.*; Vol. 3, pp.461-470.

Otter M., Årzén K., Dressler I., 2005. *TaskGraph - A Modelica Library for Hierarchical Task Machines*. Modelica 2005 Proceedings.

Seabra, E. A. R., Machado, J., Silva, J. C. L., Soares, F. O., Leão, C. P., 2007. *Simulation and Formal Verification of Real Time Systems: A Case Study*. ICINCO'2007, Angers, France; May 9-12, 6p.

Ramadge P. J. and Wonham W. M., 1987. Supervisory control of a class of discrete event processes. *SIAM J. Control Optimization*, 25(1), pp. 206-230.

Rossi O., 2004. *Validation formelle de programmes Ladder Diagram pour Automates Programmables industriels*. PhD Thesis, École Normale Supérieure de Cachan.

TOWARDS ASYNCHRONOUS SIGNAL PROCESSING

Dariusz Kościelnik and Marek Miśkowicz
Department of Electronics, AGH University of Science and Technology, al. Mickiewicza 30, Cracow, Poland
koscieln@agh.edu.pl, miskow@agh.edu.pl

Keywords: Asynchronous circuits, Analog-to-digital converters, Asynchronous Sigma-Delta modulation.

Abstract: A challenging problem of today's ADC design is a development of low voltage, low power and possibly high performance converters. The ever growing demand for decreasing the supply voltage of semiconductor devices due to scaling the feature size of VLSI technology has pushed the design of analog integrated circuit to its limits. The same problem concerns the analog-to-digital converters since lowering supply voltage results in a reduction of a voltage increment corresponding to the least significant bit (LSB) in signal amplitude quantization. In the paper, an important alternative to conventional ADCs is presented. To overcome problems with decreasing accuracy of amplitude quantization, a new class of asynchronous ADCs is discussed where the mapping of an analog signal into time domain rather than into amplitude domain is used. The asynchronous ADCs are not controlled by any global clock but self-timed. The local reference clock is used only to quantize time intervals that represent the converted signal amplitude. The design of asynchronous Sigma-Delta analog-to-digital converter (ASD-ADC) with serial output interface is discussed in details. The ASD-ADC together with the loss-free asynchronous analog signal recovery method developed recently provides possibility to establish the asynchronous digital signal processing chain.

1 INTRODUCTION

The ever growing demand for extending digital functionality on a single chip results in scaling the feature size of VLSI technology in order to increase the integration density of semiconductor devices. Scaling the CMOS transistor dimensions into nanoscale (<100 nm) enables faster operation of circuits on the one hand, but needs decreasing the supply voltage of devices to maintain reliable operation on the other. As a result of this, a design of analog and mixed signal systems has to cope with an ever increasingly challenging technological environment. For example, with the operating voltage of 1V, the output signal swing is only 0.3V, which is unacceptably low signal swing for many applications (Matsuzawa, 2007).

In the context of analog-to-digital converters (ADCs), the technology scaling increases the maximum conversion rate, but unfortunately decreases at the same time the signal-to-noise ratio (SNR). The latter is caused simply by a reduction of voltage increment corresponding to the least significant bit (LSB) in a signal amplitude quantization. This is currently the most serious problem of a classical ADC design that will be even more critical in future with further scaling of CMOS process technology feature size (e.g., in the 45 nm technology, the maximum operating voltage of around 1 V will be used).

To maintain a high SNR despite the low-voltage operation of classical ADCs, the power consumption needs to be increased (Matsuzawa, 2007). However, the latter is in general unacceptable in portable equipment and in wireless sensor networking (WSNs) due to constraints on energy resources. Efficiency of power consumption becomes a primary criterion of designing ADCs for many applications. The representative examples are environmental monitoring and biomedicine. In particular, the ADCs for WSNs in biomedical applications (pulse-oximetry, ECG, PCG, EEG, blood pressure, etc.) need only modest precision ($\leq$ 8bit), and modest speed ($\leq$ 40kHz) but has to be very energy-efficient (Yang and Sarpeshkar, 2006). Summing up, the challenging problem of today's ADC design is a development of low voltage, low power and possibly high performance ADCs whose SNR does not decrease with supply voltage reduction.

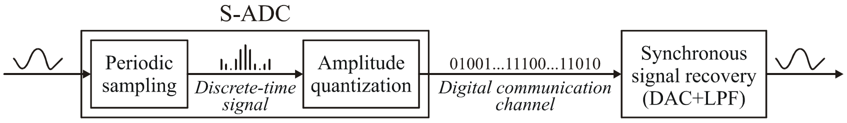


Figure 1: Traditional synchronous signal processing chain (digital numbers represent amplitude information).

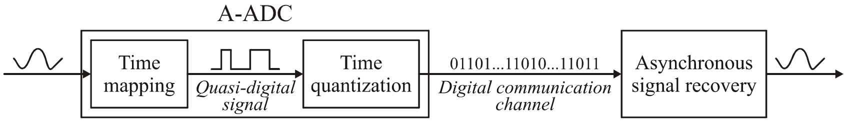


Figure 2: Asynchronous signal processing chain (digital numbers represent time information).

2 ASYNCHRONOUS ADCS

To overcome problems with decreasing accuracy of amplitude quantization, a new class of ADCs called *asynchronous analog-to-digital converters* (A-ADCs) was proposed recently where the mapping of an analog signal into *time domain* rather than into *amplitude domain* is used (Allier *et al.*, 2003). In general, the concept of time-encoding of a signal amplitude is not new since it was used for example in the well-known dual-slope ADCs, and in the frequency-to-code converters. A time-based energy-efficient ADC developed recently also uses the time as an intermediate signal variable (Yang and Sarpeshkar, 2005). In the asynchronous ADCs, the time is used as the ADC output signal so binary words that appear on the converter output irregularly represent a sequence of time intervals instead of a series of signal amplitude samples.

2.1 Sync Versus Async ADCs

The principle of *asynchronous analog-to-digital converters* (A-ADCs) is completely different from classical ADCs that are synchronous devices controlled by a global clock. In the synchronous ADCs (S-ADCs), the periodic sampling and the amplitude quantization are applied. Instead, in the A-ADCs, the analog signal is mapped into timing (quasi-digital) parameters that are further quantized according to the resolution of a reference clock. Thus, the A-ADC operation consists in a redefinition of domains at which the signal is sampled and quantizated. The A-ADC does not include sample-and-hold circuits, and is not controlled by any global clock but self-timed. The local reference clock is used only to quantize time intervals that represent the converted signal amplitude. The clockless architecture is attractive for energy-efficient design since a global clock is the primary component of power consumption in contemporary electronic instrumentation. The invention of A-ADCs announces thorough revision of the whole signal processing chain and a development of a new processing area called *asynchronous signal processing*. The synchronous and the asynchronous signal processing chains are presented in Figs. 1-2.

2.2 State of Art of A-ADCs Design

Although various techniques have been accommodated to the design of low power ADCs, the asynchronous clockless architectures based on event-based sampling have been studied in the context of ADCs in a few works only. Initially, the well-known advantages of the asynchronous design (i.e. low energy consumption, immunity to metastable behavior, reduction of the circuit average activity and electromagnetic interferences) have inspired researchers to improve operation of conventional synchronous ADCs by adoption of solutions intrinsic for the asynchronous technology. Such converters are *locally asynchronous*, but *globally synchronous* since the sampling scheme is still time-triggered and periodic (Sayiner *et al.*, 1996; Roza, 1997, Kinniment *et al.*, 2000).

A postulate of a fully asynchronous ADC based on the level-crossing sampling and the asynchronous design has been introduced by Allier et. *al* (2003). The purpose of a (fully) asynchronous ADC design is a thorough revision of the whole signal processing chain. In (Allier *et al.*, 2005), the CMOS implementation of LC-ADC with experimental results is reported. The performance index (Figure of Merit) of the LC-ADC is twice higher than that of a classical synchronous ADCs. A significant performance improvement achieved in the LC-ADC stems from reducing the activity of the asynchronous converter by a substitution of the periodic sampling by the level-crossing scheme. The average rate of the *level-crossing sampling* operations are lower than the frequency of the periodic sampling because the former are triggered if the input signal crosses prespecified levels disposed along the amplitude domain (compare Figs. 3a-3b), see (Miśkowicz, 2006) for details.

Akopyan *et al.* (2006) have designed a *level-crossing flash* ADC (LCF-ADC) dedicated to real-time monitoring and control applications where the analog signal reconstruction is not required. Instead, only the actual reports about a state of the observed object are generated. Since in the LCF-ADC the time is not tracked explicitly (i.e. the converter does not record the times at which the samples are taken), the power consumption is reduced additionally due to eliminating the circuitry that deals with time tracking. The architectures of the LCF-ADC and the LC-ADC are completely different. The latter adopts the feedback-based approach (Allier *et al.*, 2003; Allier *et al.*, 2005), and the former utilizes a parallel flash-type topology (Akopyan *et al.*, 2006).

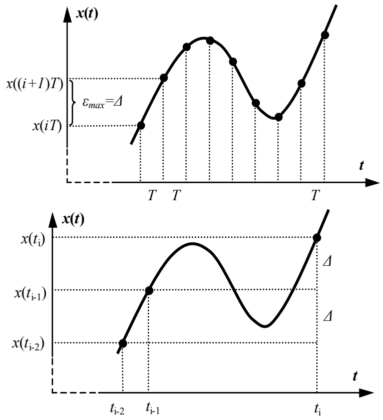


Figure 3: Comparison of the periodic (a) and the level-crossing sampling (b) schemes for the same sampling resolution, i.e. $\varepsilon_{\max} = \Delta$.

Summing up, several advantages of asynchronous ADCs in relation to conventional synchronous ADCs can be displayed as follows. The asynchronous ADCs are a low-cost alternative to conventional converters due to lower energy consumption, simple architecture, and elimination of the global clock and the sample-and-hold circuits.

2.3 Time vs. Amplitude Quantization

Although the time quantization is in general a complementary process to the amplitude quantization, certain differences might be distinguished. Whereas the analog signal amplitude is bounded and usually a non-monotone function, the time is a magnitude with unceasingly growing values. As a result, each quantization of the amplitude can be referred to a certain *absolute reference* level (usually zero). Instead, a quantization of time has to be always related to the *relative reference* which is the most recent event (i.e. a beginning of the present time interval). Next, whereas the amplitude is a fully analog magnitude, the frequency/time is considered as 'quasi-digital' domain since these parameters combine both analog and digital signal properties (Kirianaki et al., 2002). Furthermore, the time quantization is usually characterized by a non-redundant conversion time. Instead, the quantization of the amplitude takes a non-zero conversion time, sometimes is slow (e.g. in conventional successive-approximation ADCs), or the conversion time is variable and dependent on the input signal level (e.g. in delta-encoded ADCs). Finally, the frequency references (e.g. crystal oscillators) are more stable than the voltage reference sources (that are sensitive to the temperature and the technological process tolerance) so the time quantization is in general more accurate than the quantization in the amplitude domain.

3 ASD-ADC CONCEPT

In this paper, we present a concept of analog-to-digital conversion based on the asynchronous Sigma-Delta modulation. The architecture of asynchronous Sigma-Delta ADC (ASD-ADC) with serial output interface is shown in Fig. 4.

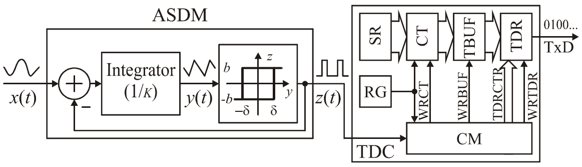


Figure 4: The architecture of ASD-ADC with serial output interface (TxD).

A two-level conversion scheme is utilized in the ASD-ADC. First, the analog signal $x(t)$ amplitude is converted to an asynchronous square wave $z(t)$ with modulated frequency and duty cycle in the asynchronous Sigma-Delta modulator (ASDM) (Fig. 5). The asynchronous square wave $z(t)$ is a *quasi-digital signal* since it is discrete in the amplitude and continuous in the time. The widths of successive pulses of $z(t)$ depend on the mean value of converted analog signal amplitude in the corresponding time windows defined by these pulses. The width of pulses has both a lower, and an upper bound.

To provide the digital output, the pulse widths are next quantized by the n-bit *time-to-digital converter* (TDC). The quantization is based on counting periods of a reference clock during each pulse of the square wave on the ASDM output. Finally, the digital code on the ASD-ADC output represents time information that encodes the analog signal amplitude. The ASD-ADC belongs to a class of the *mean value converters* that are immune to noise due to integration in the ASDM. The asynchronous serial transmitter on the ASD-ADC output transmits the digital words consisting of a number of n bits of data preceded by the *Start* bit and completed with the *Stop* bit.

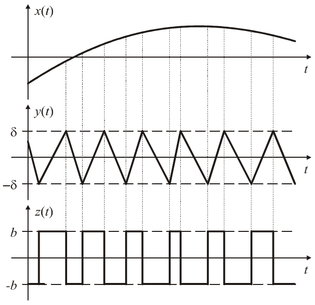


Figure 5: The waveforms on integrator output $y(t)$, and on ASDM output $z(t)$ for a given signal $x(t)$.

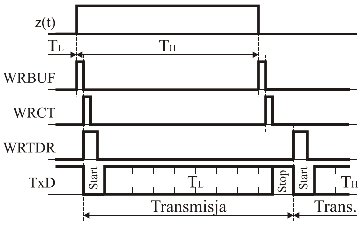


Figure 6: The timing of control signals in the time-to-digital converter (TDC).

Thus, transition times of the square wave $z(t)$ on the output of the ASDM are non-uniformly spaced. The output quantity, which is a sequence of lengths of time intervals $\Delta t_k = t_{k+1} - t_k$ between consecutive transitions, depends on the input signal behavior. The input signal $x(t)$ has to be bounded (i.e. $|x(t)| \le c$) so either the upper $\Delta t_{\max}$, or the lower bound $\Delta t_{\min}$ for Δt_k are also bounded (Lazar and Toth, 2005) as follows:

$$\Delta t_{\min} = T/[2(1+\eta)] \le \Delta t_k \le T/[2(1-\eta)] = \Delta t_{\max} \quad (1)$$

where t_k, t_{k+1} are time instants of the kth and the (k+1)th transitions, respectively, $T = 4\kappa\delta/b$ is the *self-oscillation period* (i.e. the time between consecutive slopes of $z(t)$ if the modulator is fed by the zero input signal), and $\eta = c/b$ is the *maximum modulation depth*. The κ denotes the integration constant, and the δ, b are the parameters of the Schmitt trigger (see Fig. 4). The ASDM input/output characteristics is given by (Kościelnik, Miśkowicz, 2008):

$$\Delta t_k / T = 1/\{2[1+(-1)^k \eta_k]\}\,, \quad (2)$$

where $\eta_k = \overline{x_k}/b$, $\eta_k \in (-1;1)$ is the *modulation depth in the kth time window* (t_k, t_{k+1}) of the ASDM defined as the ratio between the amplitude b on the output signal $z(t)$, and the mean signal value $\overline{x_k}$ of the input signal $x(t)$ in the time interval (t_k, t_{k+1}),

3.1 ASDM Modulator

The asynchronous Sigma-Delta modulator (ASDM) consists of the lowpass filter (integrator), and the Schmitt trigger operating in a negative feedback loop (Fig. 4). For zero input signal $x(t)$, the square wave $z(t)$ on the ASDM output oscillates with the *self-oscillation period* (T) and ½ duty cycle. The ASDM does not require any clocking and can operate at low current and supply voltage since the corresponding analog circuitry is extremely simple.

The idea of the asynchronous Sigma-Delta modulation was formulated in the 60s (Roza, 1997). However, a use of ASDM for signal conversion became especially attractive because the loss-free analog signal recovery based on ASDM output signal was developed recently (Lazar and Toth, 2005). In (Kościelnik, Miśkowicz, 2008), the ASD-ADC with the charge-pump integrator and with the single output data buffering is presented. In the present paper, we report the advanced version of the digital interface with the double data buffering providing the rate-based flow control.

3.2 LC-ADC vs. ASD-ADC

Our approach is motivated by several advantages of the proposed solution comparing to level-crossing-based ADCs (LC-ADCs) as follows.

First, in the ASD-ADC the information about the analog signal behavior is embedded *only* in a sequence of timing parameters. In other words, a digital output includes the timing information about the square wave on the output of the ASDM. Instead, in LC-ADCs, the digital data on the converter output have to include both the timing and the one-bit amplitude information about the level-crossing specification (Allier *et al.*, 2003).

Second, due to integrating input properties of the ASDM, the ASD-ADC is characterized by low susceptibility to noise making it suitable for noisy industrial environments. Instead, the LC-ADCs are sensitive to non-idealities in VLSI settings of a regular grid of amplitude levels triggering sampling operations.

Third, unlike LC-ADCs where the maximum time interval being encoded and digitized is unbounded and thus has to be arbitrary controlled by time-out, the maximum time interval in the ASD-ADC is bounded and controlled via design process.

Finally, the sampling theorem has been developed for ASD-based conversion by Lazar and Toth (2005) causing to exploit the ASD-ADC in applications where the exact recovery of original analog signal is required (e.g. audio/speech signal conversion). Thus, the ASD-based conversion supports a loss-free time-encoded signal processing.

3.3 TDC Architecture

The architecture of the *time-to-digital converter* (TDC) is shown in Fig. 4. The TDC consists of the *n*-bit counter (CT) with setup register (SR) used for programming initial states of the CT, the reference generator (RG), the control module (CM) that produces control signals for data transfer (WRCT, WRBUF, WRTDR, TDRCTR), the intermediate buffer (TBUF), and the transmitting buffer (TDR) with the serial output TxD. The timing of control signals in the TDC is shown in Fig. 6.

3.3.1 Initial State of Counter

The counting of the reference clock periods T_0 starts from an assumed initial state of the counting module defined by the number whose value is less than zero because the Δt_k is bounded by the $\Delta t_{\min}$. Thus, only the differences $\Delta t_k - \Delta t_{\min}$ might be quantized (Lazar and Toth, 2005). We have defined the optimal number M that guarantees the best resolution of the ASD-ADC. This number is negative and defined as $M = -T/4T_0$. The optimal initial number corresponds simply to the minimum pulse width that equals $T/4$ as follows from the formula (1) (Fig. 7).

3.3.2 Serial Output Interface

Unlike in conventional ADCs, the digital data appear on the ASD-ADC output irregularly according to the current variations of the analog signal amplitude. Therefore, the serial interface has to provide data flow control. The core of our concept of the TDC consists in the use of a *double data buffering* in the digital interface since the digital words appear in bursts of two words on the ASD-ADC output. This corresponds to the use of *rate-based flow control* (Verissimo, Rodrigues, 2001). The double data buffering enables overlapping a serial transmission of the *i*th digital word, a storage of the (*i*+1)th word, and a simultaneous quantization of the (*i*+2)th pulse. Thus, the serial output interface consists of two data buffers (TBUF and TDR) (Fig. 4). Each digital word obtained as result of counting is recorded and stored in the intermediate buffer (TBUF) as soon as transmission of the previous digital word is completed (Figs. 6 and 8). If so, the given digital word is transferred from the intermediate buffer (TBUF) to the transmitting buffer (TDR), which causes serial transmission of the digital word to start.

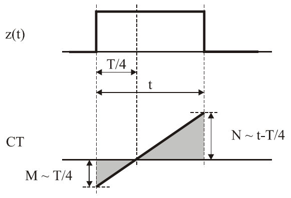


Figure 7: Counting periods of the reference clock starting from the negative initial state.

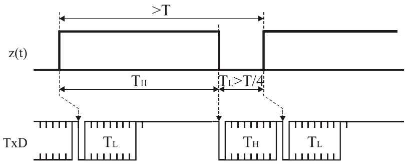


Figure 8: Serial asynchronous transmission on ASD-ADC double-buffered output port (TxD) with overlapping of transmission, storage and quantization processes (T_H denotes the digital word representing a positive pulse, and T_L represents a negative one).

3.3.3 Bit Rate on Serial Output Port

With double data buffering the minimum transmission bit rate p on the ASD-ADC serial output port is defined as: $p = 2k/T$, where k is the number of data bits of a digital word transmitted including control bits (*Start* bit, *Stop* bit, and an optional *Parity* bit).

3.3.4 Benefits of Double Data Buffering

By applying extra data buffering with the intermediate buffer TBUF, two benefits are achieved. First, the transmission bit rate on the serial output port is reduced due to shortening time intervals between successive digital words transmitted. The reduction of the transmission rate equals $(1+\eta)$ where $0<\eta<1$ is the maximum modulation depth of the ASDM so the best reduction can approach 100%. For example, with typical value $\eta=0.5$, the reduction is equal to 50%. Slowing down the transmission bit rate saves energy consumption. Second, the transmission bit rate is independent of the converted analog signal amplitude (η).

4 CONCLUSIONS

The ASD-ADC is an universal analog-to-digital converter that may be used in many applications. However, due to energy efficiency, the ASD-ADC is dedicated to use in portable devices, especially in sensors for environmental monitoring and for biomedical applications that need a long battery life. In the latter, both the wireless or skin-surface communication between sensing devices mounted on the body for health monitoring may be used (Kaldy *et al.*, 2007). In such applications, the sensors transmit data to acquisition centers at a remote side where the signals are processed, analyzed and recovered if needed. Usually, the acquisition centers access practically unlimited power. Thus, with the invention of the ASD-ADCs, energy-expensive components of signal processing chain are moved from the ADC to the locations where the energy and processing resources are available. The solution presented in the paper may be summarized as follows.

(1) The asynchronous Sigma-Delta analog-to-digital converter (ASD-ADC) together with the asynchronous analog signal recovery method (Lazar and Toth, 2005) provides possibility to establish the asynchronous digital signal processing chain where the ASD-ADC output data can be transmitted via a digital communication channel. (2) Complex and energy-expensive components of signal processing chain are moved from ADC to data acquisition center where the energy and processing resources are available. (3) The ASD-ADC digital output represents only timing information. (4) Due to higher stability of time/frequency references, the time quantization is more accurate than the voltage/current quantization. (5) Decreasing supply voltage in general does not degrade Signal-to-Noise Ratio (SNR) of the ASD-ADC. (6) With a double data buffering providing the rate-based flow control at the ASD-ADC output interface, the transmission rate is reduced even twice compared to (conventional) single-buffered interface; slowing down the transmission bit rate saves energy consumption. (7) With counting reference clock periods from the negative initial state, the dynamic range of the ASD-ADC is extended. (8) Finally, the ASD-ADC has excellent DC specification.

REFERENCES

Matsuzawa, A., 2007, Design challenges of analog-to-digital converters in nanoscale CMOS, *IEICE Trans. Electron.*, vol. E90-C, no 4, pp. 779-85.

Yang, H.Y., Sarpeshkar, R., 2006, A Bio-Inspired Ultra-Energy-Efficient Analog-to-Digital Converter for Biomedical Applications, *IEEE Trans. on Circuits and Systems-I*, vol. 53, no 11, pp. 2349-2356.

Yang, H.Y., Sarpeshkar, R., 2005, A time-based energy-efficient analog-to-digital converter, *IEEE Journal of Solid-State Circuits*, vol. 40, no 8, pp. 1590–1601.

Sayiner, N., Sorensen, H.V., Viswanathan, T.R., 1996, A level-crossing sampling scheme for A/D conversion, *IEEE Trans. on Circuits and Systems II*, vol. 43, no 4, pp. 335-9.

Roza, E., 1997, Analog-to-digital conversion via duty-cycle modulation, *IEEE Trans. on Circuits and Systems-II*, vol. 44, no 11, pp. 907–914.

Kinniment, D., Yakovlev, A., Gao B., 2000, Synchronous and asynchronous A-D conversion, *IEEE Trans. on VLSI Systems*, vol. 8, no 2, pp. 217-220.

Allier, E., Sicard, G., Fesquet L., Renaudin, M., 2003, A new class of asynchronous A/D converters based on time quantization", In *ASYNC'03*, *IEEE Intern. Symposium on Asynchronous Circuits and Systems*, pp. 196-205.

Allier, E., Goulier, J., Sicard, G., Dezzani, A., André, E., Renaudin, M., 2005, A 120nm low power asynchronous ADC, In *ISLPED 2005*, *International Symposium on Low Power Electronics and Design*, pp. 60-65.

Miśkowicz, M., 2006, Send-on-delta concept: an event-based data reporting strategy, *Sensors*, vol. 6, no 1, pp. 49-63.

Akopyan, F., Manohar, R., Apsel, A. B., 2006, A level-crossing flash asynchronous analog-to-digital converter, *In ASYNC'06, IEEE International Symposium on Asynchronous Circuits and Systems*, pp. 12-22.

Kościelnik, D., Miśkowicz, M., 2008, Asynchronous Sigma-Delta analog-to digital converter based on the charge pump integrator, *Analog Integrated Circuits & Signal Processing*, vol. 55, no 3, pp. 223–238.

Lazar, A.A., Tóth, L.T., 2005, Perfect recovery and sensitivity analysis of time encoded bandlimited signals, *IEEE Trans. on Circuits and Systems-I*, vol. 51, no 10, pp. 2060-2073.

Verissimo P., Rodrigues L., 2001, *Distributed Systems for System Architects*, Kluwer Academic Publishers.

Kaldy, C., Lazar, A.A., Simonyi, E.K., Toth, Laszlo T., 2007, Time Encoded Communications for Human Area Network Biomonitoring, Technical Report, Department of Electrical Engineering, Columbia University.

ANALOG REALIZATIONS OF FRACTIONAL-ORDER INTEGRATORS/DIFFERENTIATORS
A Comparison

Guido Maione
DEESD, Technical University of Bari, Via de Gasperi, snc, I-74100, Taranto, Italy
gmaione@poliba.it

Keywords: Non-integer-order operators, Fractional-order controllers, Rational approximation, Interlaced singularities.

Abstract: Non-integer differential or integral operators can be used to realize fractional-order controllers, which provide better performance than conventional PID controllers, especially if controlled plants are of non-integer-order. In many cases, fractional-order controllers are more flexible than PID and ensure robustness for high gain variations. This paper compares three different approaches to approximate fractional-order differentiators or integrators. Each approximation realizes a rational transfer function characterized by a sequence of interlaced minimum-phase zeros and stable poles. The frequency-domain comparison shows that best approximations have nearly the same zero-pole locations, even if they are obtained starting from different points of view.

1 INTRODUCTION

Originally, the investigation of integrals and derivatives of any order was a topic known as fractional calculus. In recent years, however, considerable attention has been paid to the concept of non-integer derivative and integral to model systems in various fields of science and engineering. In the research area of control theory, several authors have provided generalizations of classical controllers introducing various types of Fractional-Order Controllers (FOC). For example, the CRONE (French acronym for *"Commande Robuste d'Ordre Non Entièr"*) controller (Oustaloup, 1991; Oustaloup, 1995) and Fractional-Order Proportional-Integral-Derivative (FOPID) controllers $PI^{\lambda}D^{\mu}$ (Podlubny, 1999a; Podlubny, 1999b) have been recently considered. Moreover, FOC have been successfully applied in rigid robots, both for position control and for hybrid position-force-control (Tenreiro Machado and Azenha, 1998; Valerio and Sá da Costa, 2003). In general, FOC provide better performance than PID controllers, if the controlled plants are of non-integer-order. In other cases, FOC show high flexibility and can ensure high robustness for high gain variations. More particularly, in SISO systems, they can make the phase margin nearly not changing in a wide range around the gain crossover frequency, even if high gain variations produce high changes in gain crossover frequency. Applications in mechatronics are testified by several papers (Canat and Faucher, 2005; Li and Hori, 2007; Ma and Hori, 2004a; Ma and Hori, 2004b; Ma and Hori, 2007; Melchior *et al.*, 2005).

The basic element of transfer functions of FOPID controllers is the fractional differentiator/integrator s^{ν}, with ν positive or negative real number. This operator is infinite dimensional, even if it can be approximated by finite-dimension transfer functions, whose coefficients depend on the non-integer exponent ν. A good rational approximation can be obtained by truncating the continued fractions expansion (CFE) of s^{ν} (Maione, 2006; Maione, 2008). Recently, in (Barbosa *et al.*, 2006), least-squares-based methods are used for obtaining Fractional-Order Differential Filters (FODF) approximating s^{ν}.

In this paper, a novel approach is compared to two commonly used methods to realize a rational approximation of fractional-order differentiators or integrators. These operators are the basic elements in fractional-order controllers of mechatronic systems. Section 2 revisits the three different methods systematically. Section 3 compares them in the frequency domain. Section 4 draws the conclusion with some remarks.

2 REVISITING THREE RATIONAL APPROXIMATIONS

In this section, three methods are compared. They are shortly revisited, for making a direct comparison based on transfer functions putted in the same form. All the considered realizations are known to be minimum-phase and stable, with poles interlacing zeros along the negative real half-axis of the s-plane. This property is enlightened by the form of the three transfer functions, which explicitly shows the frequencies corresponding to the alternated zeros and poles. The interlacing property is important for comparison purposes, because the position of the zero-pole pairs determines the quality of the models approximating phase and magnitude of the irrational operator $(j\omega)^v$. Hence, for comparison purpose, realizations are constrained to have both their zeros with minimum module and their poles with maximum module approximately equal. All the approximating transfer functions are in a factorized form, which puts in evidence the break frequencies. Then, the lowest and highest break frequencies of the proposed method are taken as reference.

2.1 The Proposed CFE Approximation

The starting point is the following continued fractions expansion (CFE):

$$(1+x)^v \cong b_0 + \frac{a_1}{b_1 +}\frac{a_2}{b_2 +}\frac{a_j}{b_j +}\cdots \tag{1}$$

with $b_0 = b_1 = 1$, $a_1 = v\,x$ and:

$$a_j = n\,(n\text{–}v)\,x,\ b_j = 2n \tag{2}$$

$$a_{j+1} = n\,(n+v)\,x,\ b_{j+1} = 2n+1 \tag{3}$$

for $j = 2n$, with n natural number (Khovanskii, 1965). The analog approximation for the operator s^v, with $0 < v < 1$, is given in (Maione, 2008), where $x = s\text{–}1$ is used in (1) to obtain the $(2N)$-th convergent of the resulting CFE as approximating transfer functions:

$$\widetilde{G}(v,s) = \\ = \frac{p_{N0}(v)\,s^N + p_{N1}(v)\,s^{N-1} + \cdots + p_{NN}(v)}{q_{N0}(v)\,s^N + q_{N1}(v)\,s^{N-1} + \cdots + q_{NN}(v)} \tag{4}$$

where

$$p_{Nj}(v) = q_{N,N\text{-}j}(v) = \\ = (-1)^j\,C(N,j)\,(v+j+1)_{(N\text{-}j)}\,(v\text{–}\,N)_{(j)} \tag{5}$$

and $C(N,j) = \dfrac{N!}{j!(N-j)!}$ is the binomial coefficient. Moreover:

$$(v+j+1)_{(N\text{-}j)} = (v+j+1)\,(v+j+2)\,\ldots\,(v+N) \tag{6}$$

$$(v\text{–}N)_{(j)} = (v\text{–}N)\,(v\text{–}N+1)\,\ldots\,(v\text{–}N+j+1) \tag{7}$$

define the Pochammer functions with $(v\text{–}N)_{(0)} = 1$ (Spanier and Oldham, 1987). As it is easily noted, in this method the coefficients $p_{Nj}(v)$ and $q_{Nj}(v)$ are explicitly given in terms of the fractional order v. Obviously, the positions of zeros and poles in the s-plane also depend on v. So, $\widetilde{G}(v,s)$ can be written in the form:

$$\widetilde{G}(v,s) \cong \widetilde{k}\prod_{i=1}^{N}\frac{1+\dfrac{s}{\widetilde{\omega}_{z_i}}}{1+\dfrac{s}{\widetilde{\omega}_{p_i}}}\,. \tag{8}$$

As it is proved in (Maione, 2008), zeros $(-\widetilde{\omega}_{z_i})$ and poles $(-\widetilde{\omega}_{p_i})$ of $\widetilde{G}(v,s)$ are all real and interlace along the negative real half-axis in the s-plane, with:

$$\widetilde{\omega}_{z_1} < \widetilde{\omega}_{p_1} < \widetilde{\omega}_{z_2} < \widetilde{\omega}_{p_2} < \cdots < \widetilde{\omega}_{z_N} < \widetilde{\omega}_{p_N}\,. \tag{9}$$

2.2 Oustaloup's Recursive Approximation

The CRONE controller is an integer-order frequency domain approximation of s^v in the form:

$$G(v,s) \cong k\prod_{i=1}^{N}\frac{1+\dfrac{s}{\omega_{z_i}}}{1+\dfrac{s}{\omega_{p_i}}}\,. \tag{10}$$

The gain k is adjusted so that $G(v,s)$ has the same crossover frequency as the ideal operator s^v. The number N of zeros and poles of the approximating transfer function is chosen in advance. They alternate on the negative real half-axis of the s-plane so that the frequencies satisfy:

$$\omega_{z_1} < \omega_{p_1} < \omega_{z_2} < \omega_{p_2} < \cdots < \omega_{z_N} < \omega_{p_N}\,. \tag{11}$$

In this way, zeros and poles interlace on the negative real half-axis, leading to a gain which is, approximately, a linear function of the logarithm of frequency. The phase is nearly constant and approximates $\nu\,\pi\,/\,2$. The parameters ω_{z_i} and ω_{p_i} are determined by placing zeros and poles as follows:

$$\alpha = \left(\frac{\omega_H}{\omega_L}\right)^{\frac{\nu}{N}};\ \eta = \left(\frac{\omega_H}{\omega_L}\right)^{\frac{1-\nu}{N}};\ \omega_{z_1} = \omega_L\sqrt{\eta} \tag{12}$$

$$\omega_{p_i} = \omega_{z_i}\,\alpha \qquad i = 1, ..., N \tag{13}$$

$$\omega_{z_{i+1}} = \omega_{p_i}\,\eta \qquad i = 1, ..., N\text{–}1. \tag{14}$$

The frequencies ω_L and ω_H are appropriately chosen as $\omega_L < \tilde{\omega}_{z_1}$ and $\omega_H > \tilde{\omega}_{p_N}$, so that it holds $\omega_{z_1} \cong \tilde{\omega}_{z_1}$ and $\omega_{p_N} \cong \tilde{\omega}_{p_N}$.

2.3 Matsuda's Approximation

The Matsuda's method approximates the operator s^ν from its gain ω^ν. The gain is determined at $2N$+1 frequencies ω_0, ω_1, …, ω_{2N}, which are taken logarithmically spaced in the approximation interval. The interval $[\omega_0,\ \omega_{2N}]$ is chosen so that the lowest break frequency $\hat{\omega}_{z_1}$ and the highest break frequency $\hat{\omega}_{p_N}$ in the model satisfy: $\hat{\omega}_{z_1} \cong \tilde{\omega}_{z_1}$ and $\hat{\omega}_{p_N} \cong \tilde{\omega}_{p_N}$, respectively. Note that, usually, an odd value of N is used, so that the resulting approximation is proper. Then, the following functions are defined:

$$\begin{aligned} &m_0(\omega) = \omega^\nu;\ m_1(\omega) = \frac{\omega - \omega_0}{m_0(\omega) - m_0(\omega_0)};\ldots \\ &\ldots;\ m_k(\omega) = \frac{\omega - \omega_{k-1}}{m_{k-1}(\omega) - m_{k-1}(\omega_{k-1})};\ldots \\ &m_{2N}(\omega) = \frac{\omega - \omega_{2N-1}}{m_{2N-1}(\omega) - m_{2N-1}(\omega_{2N-1})} \end{aligned} \tag{15}$$

from which the following set of parameters are obtained:

$$\alpha_0 = (\omega_0)^\nu \tag{16}$$

$$\alpha_k = \frac{\omega_k - \omega_{k-1}}{m_{k-1}(\omega_k) - m_{k-1}(\omega_{k-1})} \tag{17}$$

for k = 1, 2, …, 2N.

Using the ω_k and α_k, the CFE can be written as:

$$s^\nu \cong \alpha_0 + \frac{s-\omega_0}{\alpha_1 +}\,\frac{s-\omega_1}{\alpha_2 +}\,\frac{s-\omega_2}{\alpha_3 +}\cdots \tag{18}$$

whose convergents provide the rational approximations to the irrational operator s^ν. The (2N)-th convergent of (18) can be easily converted into the rational approximation, as the ratio $\hat{G}(\nu,s)$ of two polynomials with degree N. Then, the factorization of these polynomials leads to:

$$\hat{G}(\nu,s) \cong \hat{k}\prod_{i=1}^{N}\frac{1+\dfrac{s}{\hat{\omega}_{z_i}}}{1+\dfrac{s}{\hat{\omega}_{p_i}}}. \tag{19}$$

Numerical experiments show that, also in this case, it holds:

$$\hat{\omega}_{z_1} < \hat{\omega}_{p_1} < \hat{\omega}_{z_2} < \hat{\omega}_{p_2} < \cdots < \hat{\omega}_{z_N} < \hat{\omega}_{p_N}. \tag{20}$$

3 A COMPARISON BETWEEN THREE METHODS

The approaches of the previous sections are here compared, by choosing N = 3 and then N = 4. These values are chosen to make the order of the FOC realizations as low as possible, compatibly with good performances. Figures 1, 2, 3 and 4 show the Bode plots of phase and amplitude, for the typical fractional order ν = 0.5. Other values of the integer N and of ν, with $0 < \nu < 1$, can be considered. As previously stated, the approximation is performed so that $\tilde{G}(\nu,s)$, $G(\nu,s)$ and $\hat{G}(\nu,s)$ have their first zero-frequency and their last pole-frequency nearly equal. Hence, the zero-frequency $\tilde{\omega}_{z_1}$ and the pole-frequency $\tilde{\omega}_{p_3}$ or $\tilde{\omega}_{p_4}$ of $\tilde{G}(\nu,s)$ are assumed as reference. In conclusion, it must nearly hold: $\omega_{z_1} \cong \tilde{\omega}_{z_1}$, $\hat{\omega}_{z_1} \cong \tilde{\omega}_{z_1}$, $\omega_{p_3} \cong \tilde{\omega}_{p_3}$, and $\hat{\omega}_{p_3} \cong \tilde{\omega}_{p_3}$, when N = 3, and $\omega_{z_1} \cong \tilde{\omega}_{z_1}$, $\hat{\omega}_{z_1} \cong \tilde{\omega}_{z_1}$, $\omega_{p_4} \cong \tilde{\omega}_{p_4}$, and $\hat{\omega}_{p_4} \cong \tilde{\omega}_{p_4}$, when N = 4.

First, the parameters of $\tilde{G}(\nu,s)$ are determined. For ν = 0.5 and N = 3, formula (8) gives: $\tilde{\omega}_{z_1} = 0.0521$, $\tilde{\omega}_{z_2} = 0.6360$, $\tilde{\omega}_{z_3} = 4.3119$,

$\tilde{\omega}_{p_1} = 0.2319$, $\tilde{\omega}_{p_2} = 1.5724$, $\tilde{\omega}_{p_3} = 19.1957$, and $\tilde{k} = 0.1429$. These values clearly indicate that $\tilde{G}(\nu, s)$ is minimum-phase, stable, with interlacing zeros and poles. Figure 1 reports the phase Bode diagram of $arg[\tilde{G}(\nu, j\omega)]$ (Maione's curve).

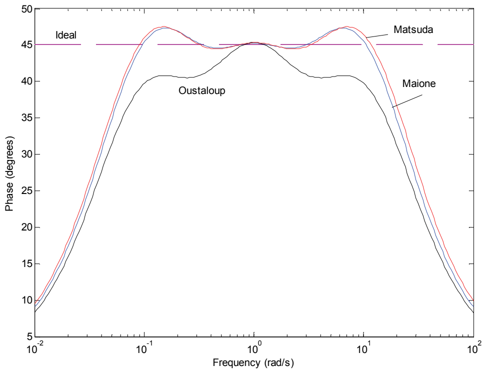


Figure 1: Phase Bode diagram for the approximations of order 3 a fractional-order differentiator, $\nu = 0.5$.

Now, the procedure for determining the function $G(\nu,s)$ is considered. With reference to (12), the interval $[\omega_L, \omega_H]$ is chosen larger than $[\tilde{\omega}_{z_1}, \tilde{\omega}_{p_3}]$. More precisely, $\omega_L = \tilde{\omega}_{z_1} \lambda_1$ and $\omega_H = \tilde{\omega}_{p_3} \lambda_2$, where λ_1 and λ_2 are coefficients to be fixed so that the Oustaloup's algorithm leads to $\omega_{z_1} \cong \tilde{\omega}_{z_1}$ and $\omega_{p_3} \cong \tilde{\omega}_{p_3}$. These coefficients are chosen by a rule of thumb. Since $\tilde{\omega}_{z_1} = 0.0521$ and $\tilde{\omega}_{p_3} = 19.1957$, simple computer experiments in MATLAB® show that choosing $\lambda_1 = 0.55$ and $\lambda_2 = 1.8$ yields: $\omega_{z_1} = 0.0518$, $\omega_{z_2} = 0.5509$, $\omega_{z_3} = 5.8634$, $\omega_{p_1} = 0.1688$, $\omega_{p_2} = 1.7972$, $\omega_{p_3} = 19.1293$, $k =$ 0.1692. As it is noted, the constraints $\omega_{z_1} \cong \tilde{\omega}_{z_1}$ and $\omega_{p_3} \cong \tilde{\omega}_{p_3}$ are respected. In Figure 1, $arg[G(\nu, j\omega)]$ is also reported (Oustaloup's curve).

Finally, for applying the Matsuda's method, the sampling frequencies are logarithmically distributed inside the approximation interval, so that it must result: $\hat{\omega}_{z_1} \cong \tilde{\omega}_{z_1}$ and $\hat{\omega}_{p_3} \cong \tilde{\omega}_{p_3}$, as requested. This result is achieved by choosing $\omega_{2N} = \lambda \tilde{\omega}_{z_1}$ and $\omega_0 = \tilde{\omega}_{p_3} / \lambda$. The parameter λ is fixed by computer experiments to $\lambda = 45$. Namely, the following breaking frequencies of $\hat{G}(\nu, j\omega)$ result: $\hat{\omega}_{z_1} = 0.0485$, $\hat{\omega}_{z_2} = 0.6248$, $\hat{\omega}_{z_3} = 4.5311$, $\hat{\omega}_{p_1} = 0.2207$, $\hat{\omega}_{p_2} = 1.6004$, $\hat{\omega}_{p_3} = 2\,0.6273$, and $\hat{k} = 0.1373$. These values show that the constraints $\hat{\omega}_{z_1} \cong \tilde{\omega}_{z_1}$ and $\hat{\omega}_{p_3} \cong \tilde{\omega}_{p_3}$ are also satisfied. As it can be easily observed, however, all the remaining frequencies and the gain of the Matsuda's model are nearly equal to those of the author's approximating transfer function. This fact is confirmed by the behaviour of $arg[\hat{G}(\nu, j\omega)]$ in Figure 1 (Matsuda's curve). The Bode plot, indeed, is nearly indistinguishable from the plot of $arg[\tilde{G}(\nu, j\omega)]$.

In conclusion, Figure 1 shows that $arg[\hat{G}(\nu, j\omega)]$ and $arg[\tilde{G}(\nu, j\omega)]$ are nearly flat and give a good approximation of $arg[(j\omega)^{\nu}] = \nu\,\pi / 2$. The plot of $arg[\hat{G}(\nu, j\omega)]$ yields a slightly worst approximation. Figure 2 confirms that the magnitude plots of $\hat{G}(\nu, s)$ and $\tilde{G}(\nu, s)$ are nearly coincident. They give a better approximation of ω^{ν} than $G(\nu,s)$, also in this case.

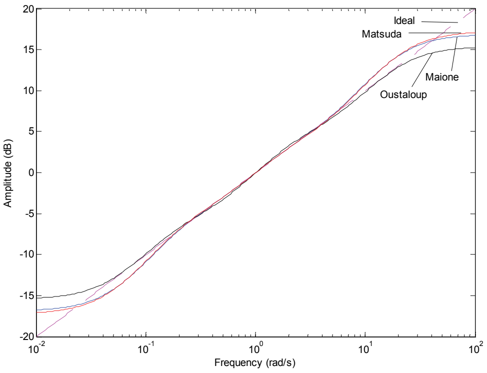


Figure 2: Amplitude Bode diagram for the approximations of order 3 of a fractional-order differentiator, $\nu = 0.5$.

Now, let us consider a different approximation obtained by using $N = 4$ and the same procedure.

For $\nu = 0.5$, formula (8) gives: $\tilde{\omega}_{z_1} = 0.0311$, $\tilde{\omega}_{z_2} = 0.3333$, $\tilde{\omega}_{z_3} = 1.4203$, $\tilde{\omega}_{z_4} = 7.5486$, $\tilde{\omega}_{p_1} = 0.1325$, $\tilde{\omega}_{p_2} = 0.7041$, $\tilde{\omega}_{p_3} = 3.0000$, $\tilde{\omega}_{p_3} = 32.1634$, and $\tilde{k} = 0.1111$. Then, $\tilde{G}(\nu, s)$ is minimum-phase, stable, with interlacing zeros and poles. Figure 3 shows the phase Bode diagram of $arg[\tilde{G}(\nu, j\omega)]$ (Maione's curve).

For the Oustaloup's approximation, $\lambda_1 = 0.61$ and $\lambda_2 = 1.64$ yield: $\omega_{z_1} = 0.0311$, $\omega_{z_2} = 0.2261$, $\omega_{z_3} = 1.6419$, $\omega_{z_4} = 11.9237$, $\omega_{p_1} = 0.0839$, $\omega_{p_2} = 0.6093$, $\omega_{p_3} = 4.4247$, $\omega_{p_3} = 32.1323$, and $k = 0.1377$. The constraints $\omega_{z_1} \cong \tilde{\omega}_{z_1}$ and

$\omega_{p4} \cong \tilde{\omega}_{p4}$ are respected. In Figure 3, $arg[G(\nu, j\omega)]$ is also reported (Oustaloup's curve).

For the Matsuda's approximation, λ = 39 gives: $\hat{\omega}_{z_1} = 0.0310$, $\hat{\omega}_{z_2} = 0.3327$, $\hat{\omega}_{z_3} = 1.4211$, $\hat{\omega}_{z_4} = 7.5702$, $\hat{\omega}_{p_1} = 0.1321$, $\hat{\omega}_{p_2} = 0.7035$, $\hat{\omega}_{p_3} = 3.0055$, $\hat{\omega}_{p_4} = 32.2772$, and $\hat{k} = 0.1109$. For N = 4, the frequency response of $arg[\hat{G}(\nu, j\omega)]$ is practically indistinguishable from that of $arg[\tilde{G}(\nu, j\omega)]$ (Matsuda's and Maione's curves are practically the same).

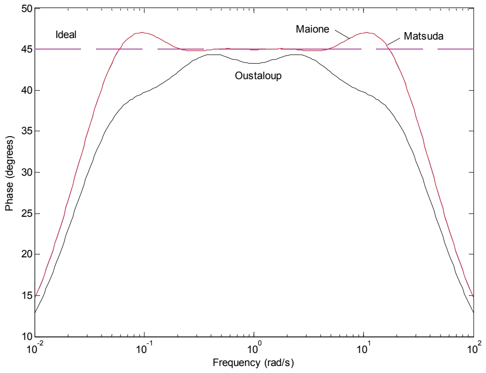


Figure 3: Phase Bode diagram for the approximations of order 4 of a fractional-order differentiator, ν = 0.5.

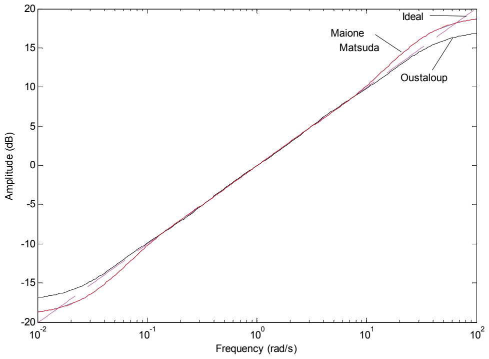


Figure 4: Amplitude Bode diagram for the approximations of order 4 of a fractional-order differentiator, ν = 0.5.

Figure 4 confirms that the magnitude plots of $\hat{G}(\nu, s)$ and $\tilde{G}(\nu, s)$ are nearly the same and give a better approximation of ω^{ν} than $G(\nu,s)$, for N = 4.

4 CONCLUDING REMARKS

This paper compared three different methods to approximate non-integer-order differential or integral operators in fractional-order controllers: these methods are the author's, the Oustaloup's, and the Matsuda's, respectively. All approximations of the irrational operator s^{ν} were realized through analog transfer functions characterized by stable poles and minimum-phase zeros. In particular, zeros and poles were interlaced along the negative real half-axis of the s-plane, and the first and last singularities were constrained to be nearly the same in all approximations. The interlacing property allowed us the comparison to find the best distribution of singularities. Namely, a frequency domain analysis of the phase diagrams showed that the author's and Matsuda's approximations outperformed the well-known by Oustaloup.

Note that all realizations were limited to the lowest order that could guarantee good performance. The better results achieved by the proposed approximation are due to a better distribution of interlaced zeros and poles. It is also interesting to note how the proposed approximation achieves nearly the same zero-pole pairs of the Matsuda's approximation, even if the starting points of the two methods are completely different.

REFERENCES

Barbosa, R.S., Tenreiro Machado, J.A., Silva, M.F., 2006. Time domain design of fractional differintegrators using least-squares. *Signal Processing*, Vol. 86, No. 10, pp. 2567-2581.

Canat, S., Faucher, J., 2005. Modeling, identification and simulation of induction machine with fractional derivative. In *Fractional Differentiation and its Applications*, Le Mehauté, A., Tenreiro Machado, J.A., Trigeassou, J.C., Sabatier, J. (Eds.), Ubooks Verlag Ed., Neusäß, Vol. 2, pp. 195-206.

Khovanskii, A.N., 1965. Continued fractions. In Lyusternik, L.A., Yanpol'skii, A.R. (Eds.): *Mathematical Analysis - Functions, Limits, Series, Continued Fractions*, chap. V, Pergamon Press. Oxford, International Series Monographs in Pure and Applied Mathematics (transl. by D. E. Brown).

Li, W., Hori, Y., 2007. Vibration suppression using single neuron-based PI fuzzy controller and fractional-order disturbance observer. *IEEE Transactions on Industrial Electronics*, Vol. 54, No. 1, pp. 117-126.

Ma, C., Hori, Y., 2004a. Backlash vibration suppression control of torsional system by novel fractional-order PID^{k} controller. *Transactions of IEE Japan on Industry Application*, Vol. 124, No. 3, pp. 312-317.

Ma, C., Hori, Y., 2004b. Fractional order control and its application of PI$^{\alpha}$D controller for robust two-inertia speed control. In *Proceedings of the 4th International Power Electronics and Motion Control Conference*

(IPEMC 04), Xi'an, China, 14-16 Aug. 2004, Vol. 3, pp. 1477-1482.

Ma, C., Hori, Y., 2007. Fractional-order control: theory and applications in motion control (past and present). *IEEE Industrial Electronics Magazine*, Winter 2007, Vol. 1, No. 4, pp. 6-16.

Maione, G., 2006. Concerning continued fractions representation of noninteger order digital differentiators. *IEEE Signal Processing Letters*, Vol. 13, No. 12, pp. 725-728.

Maione, G., 2008. Continued fractions approximation of the impulse response of fractional order dynamic systems. *IET Control Theory & Applications*, Vol. 2, No. 7, pp. 564-572.

Melchior, P., Sabatier, J., Duboy, D., Ferragne, H., Amagat, C., 2005. CRONE position controller for pneumatic butterfly valve controller by on-off valves. In *Fractional Differentiation and its Applications*, Le Mehauté, A., Tenreiro Machado, J.A., Trigeassou, J.C., Sabatier, J. (Eds.), Ubooks Verlag Ed., Neusäß, Vol. 3, Chap. 16, pp. 721-734.

Oustaloup, A., 1991. *La Commande CRONE. Commande Robuste d'Ordre Non Entièr*, Editions Hermès. Paris, France.

Oustaloup, A., 1995. *La Dérivation non Entière: Théorie, Synthèse et Applycations*, Editions Hermès, Serie Automatique. Paris, France.

Podlubny, I., 1999a. *Fractional Differential Equations*, Academic Press. San Diego, CA, USA.

Podlubny, I., 1999b. Fractional-order systems and $PI^{\lambda}D^{\mu}$ controllers. *IEEE Transactions on Automatic Control*, Vol. 44, No. 1, pp. 208-214.

Spanier, J., Oldham, K.B., 1987. *An atlas of functions*, Hemisphere Publishing Co.. New York, 1987.

Tenreiro Machado, J.A., Azenha, A., 1998. Fractional-order hybrid control of robot manipulators. In *IEEE SMC'98, Proceedings of the 1998 IEEE International Conference on Systems, Man, and Cybernetics*, Hyatt La Jolla, San Diego (CA), USA, 11-14 Oct. 1998, pp. 788-793.

Valerio D., Sá da Costa, J., 2003. Digital implementation of non-integer control and its application to a two link control arm. In *Proceedings of the European Control Conference*, Cambridge, UK, 1-4 Sept. 2003.

LINEAR IDENTIFICATION OF ROTARY WHITE CEMENT KILN

Golamreza Noshirvani, Mansour Shirvani
ACECR Markazi Branch, Beheshti Street, Arak, Iran
rnoshirvani@jdmarkazi.ac.ir, Shirvani.m@iust.ac.ir

Alireza Fatehi
Control Departmen , KNtoosi University, Tehran, Iran
fatehi@kntu.ac.ir

Keywords: Rotary White Cement Kiln, Dynamic Modelling, Linear Identification, Box-Jenkins.

Abstract: Rotary cement kiln is the main part of a cement plant that clinker is produced in it. Continual and prolonged operation of rotary cement kiln is vital in cement factories. However, continual operation of the kiln is not possible and periodic repairs of the refractory lining would become necessary, due to non-linear phenomena existing in the kiln, such as sudden falls of coatings in the burning zone and probability of damages to the refractory materials during production. This is the basic reasoning behind the needs for a comprehensive model which is severely necessary for better control of this process. Such a model can be derived based on the mathematical analysis with consultation of expert operator experiences. In this paper linear model is identified for rotary kiln of Saveh white cement factory. The linear model is introduced using Box-Jenkins structure. The results of the obtained model were satisfactory compared to some other models and can be used for designing adaptive or robust controllers.

1 INTRODUCTION

During the years of clinker production, many changes and improvements have been occurred. Rotary kilns is not just for cement production, while it is used in different chemical industries such as lime burning, crude oil calcinations, solid garbage ash, titanium dioxide calcinations, aluminium oxide process and etc. In all cases, use of rotary kiln, due to its basic role about energy consumption, desired reaction performance and many other advantages is preferred. However, control of the kiln in optimal condition is of primary importance and is not possible unless having a good knowledge and a comprehensive model based on important phenomena occurring in the system. In this way, several research papers have been published among which the original modelling of Spang, based on material and energy balance of the kiln can be mentioned (Spang, 1972). In Spang's mechanistically model several assumptions have been used. Also, Frish assumed the kiln as cylindrical vessel with internal non adiabatic heating source, and focused on the monotonically state. He assumed the heat transfer based on radiated rather than displacement type (Frish and Jeschar, 1983). In figure 1, a schematic of the kiln with its cyclone pre-heater is shown.

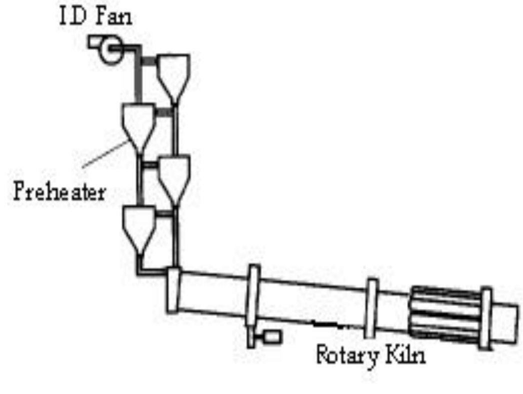


Figure 1: Rotary Cement kiln Process.

In this paper we will use a black box identification procedure for modelling the Saveh white cement kiln. It is a 65 m long, 4.7 m diameter kiln with 4 stage double string pre-heater and water immersion cooler. The main manipulated variables of the kiln are:

- kiln speed
- Fuel Flow Rate
- ID. Fan speed
- Raw Material Flow Rate

Also the output variables according to the operator's experiences are as following:

- back end temperature
- Remained (unused) oxygen, O_2
- CO content of outlet gases from the kiln
- Kiln DC motor current
- Preheater temperature
- Cooler temperature

Based on these variables the rotary kiln is a 4-inputs 6-outputs plant as shown in figure 2.

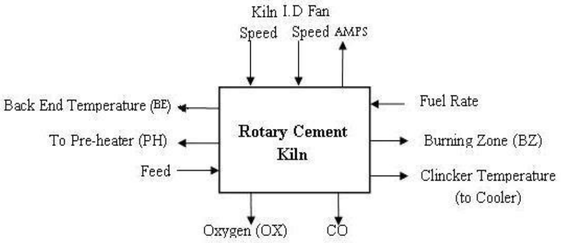


Figure 2: Block diagram of the kiln based on input-output.

These variables are so important and selected with fundament of expert operators, such as with these input, the kiln can be controlled. The burning zone temperature is not only one of the most important kiln control variables but also the most difficult one to monitor (Peray, 1986). Despite the fact that burning zone condition in modern kilns are shown as temperature profile that used for manual controllers.

2 CONDITION OF DATA GATHERING

There are three important factors in modelling based on the identification techniques:

- Useful and valid data
- A perfect and useful model
- Strong method to adjust the model

Input and outputs must be selected such that input change affects output variables. Also the recognition of process behavior will be much simpler if input-output data is reach, i.e. it consists different operating points and frequency contents. However, system identification based on input output data does not introduce a physical model with exact structure but it does a model that fits the data. Therefore, selection of a proper model is important. Also, the obtained data should have the process information to be used for identification.

An important point concerning data gathering is that to be careful that the disturbances and unexpected events such as creation of coating and coating fall in the kiln and do not change the system behaviour. The white cement rotary kiln identification is passive process, meaning that we can only observe the plant variables under a given circumstance and it is technically impossible to introduce extra excitation on these systems. The data from these systems may not be informative enough. This can make the identification of the system difficult (Zhu, 2001). Therefore it is not possible to expect from the presented model to have the same behaviour with the real system in an abnormal condition unless these conditions are occurred a few times during data gathering.

For this reason, data gathered during a period of 18 hours for several times. Finally, the best conditioned data were obtained for the rotary kiln in 2008-05-07. Figure 3 shows the input variables. The output variables are shown in figure 4.

3 DATA PRETREATMENT

After collecting perfect data from rotary kiln, the data will not be used directly for identification process. One of its reasons is high frequency noises and spikes on the main signals. Sometimes immeasurable disturbances occur and take the system out of its linear range. Changing Operation point causes entering nonlinear effects in output data. To solve the problem of high frequency noises and some of these problems, it is tried to use some pre-processing methods mentioned in identification references to reach a perfect model of process (Ljung, 1999; Nelles, 2001) . For considering rest of them, we tried to choose the model structure and focus on its flexibilities.

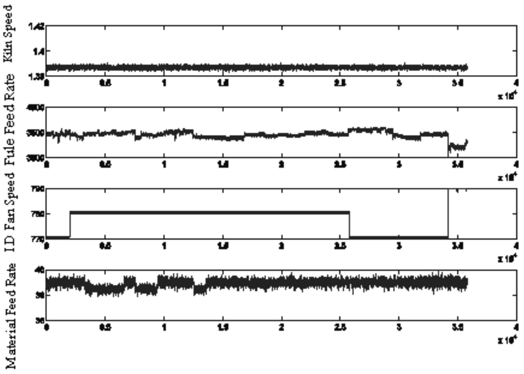


Figure 3: Input data representation for white cement kiln identification.

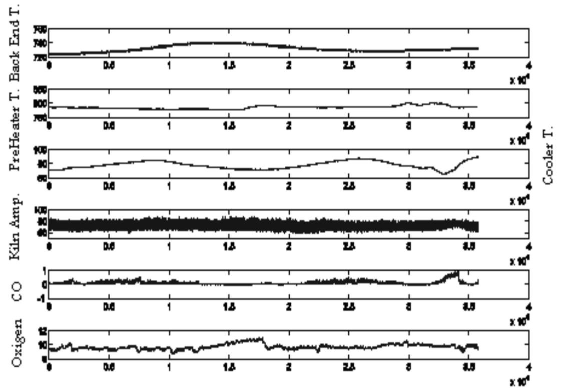


Figure 4: Output data representation for white cement kiln identification.

3.1 Peak Shaving

At the first stage, it is necessary to pay attention on data to recognize the system dynamics based on the available input and output data. It is important to smooth the spikes and shave the peaks. Spikes are because of sensors operation or data acquisition card that causes a numerical fault in data representation (Astrom, 1984), whereas the high energy of spikes interfere the model parameters estimation and its validation. Applying a third order digital Butterworth filter on the data gathered from the kiln.

The filtered output for kiln back temperature is shown in figure 5. Correlation analysis is used to obtain the weight and important dynamics between input and output data (Noshirvani, 2005). The similarity of two signals will be measured in correlation analysis. In this analysis, the correlation order of two signals is measurable. These contexts can be written as the following formulas:

$$\Phi_{u,y}(\tau)=E\{u(t)\cdot y(t-\tau)\}=\lim_{N\to\infty}\frac{1}{N}\sum_{k=1}^{N}u(k)\cdot y(k-\tau) \qquad (1)$$

where $\Phi_{u,y}(\tau)$ is the cross correlation of u and y and

$$\Phi_{u,u}(\tau)=E\{u(t)\cdot u(t-\tau)\}=\lim_{N\to\infty}\frac{1}{N}\sum_{k=1}^{N}u(k)\cdot u(k-\tau) \qquad (2)$$

where $\Phi_{u,u}(\tau)$ is the autocorrelation of u.

Correlation analysis assumes a linear system and does not require a specific model structure; also it could be used to assess the effective dynamics.

Figure 6 shows the correlation between input fuel rate and the kiln speed to burning zone temperature.

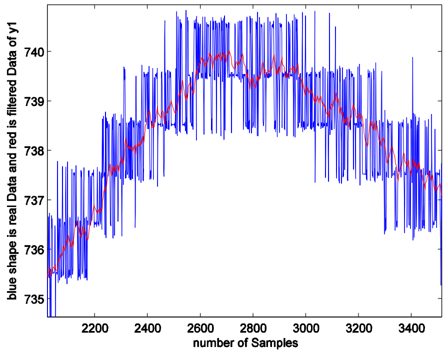


Figure 5: Real data and Filtered data representation.

The basic assumption in the discussion is that the identification model will be used in control. Therefore the main dynamics used for this output in identification have been shown, and then the plant is broken into 6 MISO models.

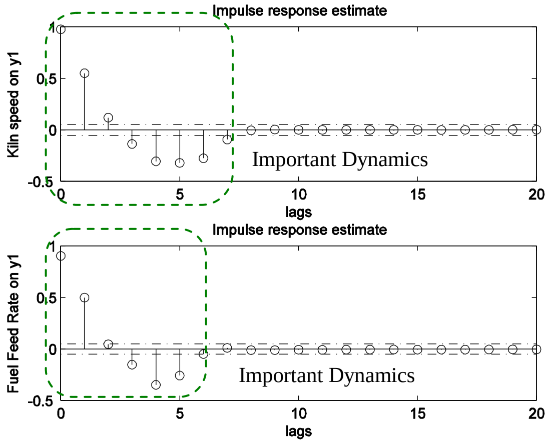


Figure 6: Correlation between first and second inputs with the first output.

4 LINEAR IDENTIFICATION

Different linear models were studied for system identification. The best obtained linear model for the kiln was Box-Jenkins (BJ) model which its result is explained here. BJ model is defined as:

$$y(k)=\frac{B(q)}{F(q)}u(k)+\frac{C(q)}{D(q)}\upsilon(k) \qquad (3)$$

This structure has been introduced by Box-Jenkins in 1970. The predictor for this model is illustrated in (5).

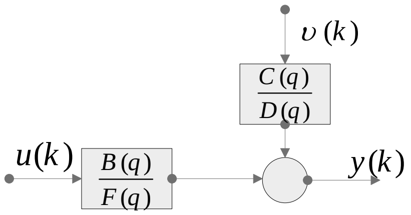


Figure 7: Box-Jenkins Structure.

$$\begin{aligned} F(q) &= 1+f_1q^{-1}+\cdots+f_nq^{-n} \\ B(q) &= b_1q^{-1}+\cdots+b_nq^{-n} \\ C(q) &= 1+c_1q^{-1}+\cdots+c_nq^{-n} \\ D(q) &= 1+d_1q^{-1}+\cdots+d_nq^{-n} \end{aligned} \tag{4}$$

$$\hat{y}(k\,|k-1) = \frac{D(q)B(q)}{C(q)F(q)}u(k) + \left[1-\frac{D(q)}{C(q)}\right]y(k) \tag{5}$$

where $\hat{y}(k)$ is the output of the model. The notation |*k-1* is used because the optimal prediction of Box-Jenkins model utilizes previous process outputs in order to extract the information contained in the correlated disturbance, *n(k)* affects on output variable, that is defined in (6) and the prediction of error of this model can be obtained with (7).

$$e(k) = \frac{D(q)}{C(q)}y(k) - \frac{B(q)D(q)}{F(q)C(q)}u(k) \tag{6}$$

Box-Jenkins model is estimated by nonlinear optimization, where first an auto regressive estimation to determine the initial parameter values for b_i and f_i. The gradients of models function can be computed as follows.

$$\begin{aligned} &F(q)C(q)\hat{y}(k\,|k-1) = B(q)D(q)u(k) + \\ &F(q)\;\; C(q) - D(q)\;\; y(k) \end{aligned} \tag{7}$$

Differentiation of (7) with respect to b_i yields

$$F(q)C(q)\frac{\partial\hat{y}(k\,|k-1)}{\partial b_i} = D(q)u(k-i) \tag{8}$$

This leads to

$$\frac{\partial\hat{y}(k\,|k-1)}{\partial b_i} = \frac{D(q)}{F(q)C(q)}u(k-i) \tag{9}$$

Also these computations have done for c_i, d_i and f_i. The parameters of this model will be trained based on minimizing of the following cost function:

$$V_{BJ} = \frac{1}{N}\sum_{t=1}^{N}\left[\frac{D(q)}{C(q)}\left\{y(t) - \frac{B(q)}{F(q)}u(t)\right\}\right]^2 \tag{10}$$

The main advantage of Box-Jenkins is giving a better estimation for the closed-loop models, but its implementation is a challenging task (Eykoff, 1974).

In general Box-Jenkins (BJ) model has several advantages over the output error method. Firstly, it will supply both a process model and a disturbance model. As shown in table1, this model will be consistent also in passive identification; this implies that this method will give a more accurate process model than an output error method for a given process under passive data condition. However, the BJ model has a more complex structure, which implies that numerical optimization will be more complicated.

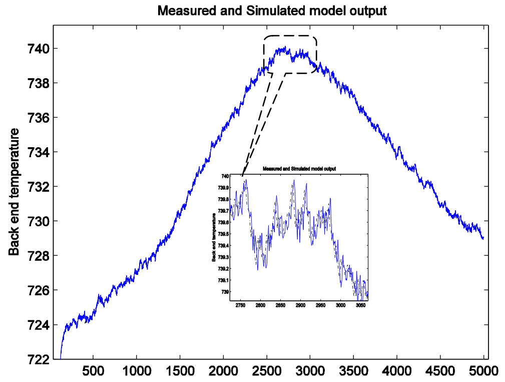


Figure 8: Actual and simulated signal of kiln back-end temperature.

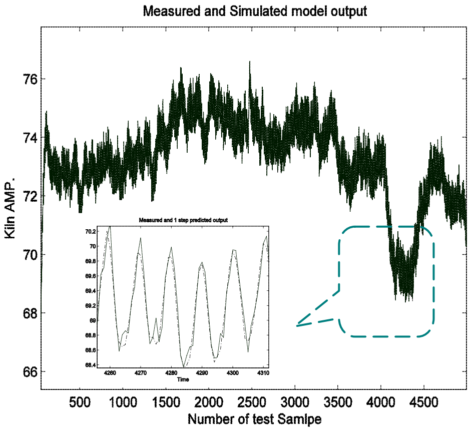


Figure 9: Actual and simulated signal of Kiln motor current.

Carbon monoxide analysis, because it samples dirty kiln gases and takes the sample at a location where high temperature prevail, has a tendency to multifunction frequently unless almost daily preventive maintenance is carried out on this unit. The location, where the sample probe is installed, is also key point to consider as false air in leakage could distort the true contents of CO in the exit gases (Shirvani et.al, 2004).

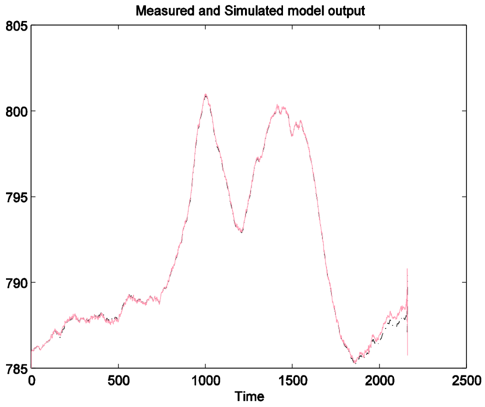


Figure 10: Actual and simulated signal of Pre-heater temperature.

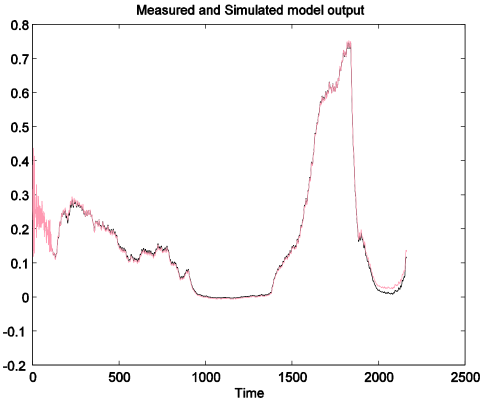


Figure 11: Actual and simulated signal of Carbon monoxide.

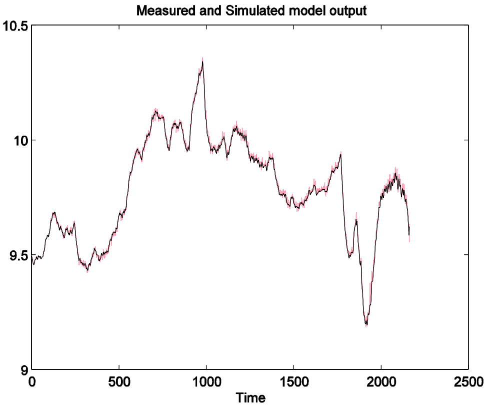


Figure 12: Actual and simulated signal of Oxygen.

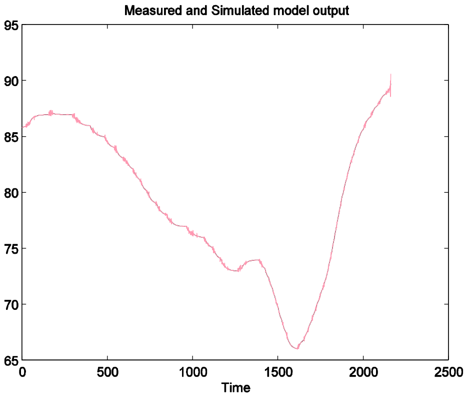


Figure 13: Actual and simulated signal of Cooler temperature.

Equation11 is known as the mean square error of model. It is an estimation of the standard deviation of the model error with respect to data.

$$E = \frac{1}{N}\sum_{K=1}^{N}\left|y(k)-\hat{y}(k)\right|^2 \qquad (11)$$

The models are compared with test data that obtained also in 2008-08-15 for # hours from control system of Saveh white cement plant. The fitness of the model with the plant can be computed as (Eykoff, 1974). Then the best criterion is (12).

$$fitness = \left(1-\frac{\left|y-\hat{y}\right|}{\left|y-\bar{y}\right|}\right)\times 100 \qquad (12)$$

where $\bar{y}(k)$ is the mean of $y(k)$.

By comparing different dynamic models like output error (OE) and ARMAX in equation (12) can be concluded that BJ modelling has a better response (Noshirvani, 2005). This result is because passive modelling of kiln system is severely non-linear. Therefore, as it is shown in Table1, the most enriched linear model has relatively better performance.

Table 1: comparing different linear Models of plant.

Variable	B.J.	ARMAX	O.E.
Back end temperature	85%	64%	37%
Current motor Kiln	84%	70%	29%
Preheater Temperature	91%	75%	40%
Carbon mono oxide	88%	68%	39%
Oxygen	89%	68%	30%
Cooler temperature	87%	61%	35%

5 CONCLUSIONS

In this paper, some linear approaches for system identification and model parameter estimation have been applied to an industrial scale white cement kiln.

Since the white cement rotary kiln identification is passive and the process input data were inadequate and the signal to noise rate was very high, it is a complex process which needs some comprehensive identification procedure. Different kind of linear models are examined in which BJ dynamic model presents the best result compare to other linear models.

Linear structure can be used for identifying the rotary cement kilns, but in this procedure, the coating fall and its creation in the kiln, will be ignored. Thus, the train and test data have been gathered with this assumption.

Weakness of linear modelling based on O.E is that it is proper for slow damping process, but in this plant slow dynamics related to the system and fast dynamics related to noise are not completely segregated and obtained error is mostly related to enforced noise in output signals.

REFERENCES

Spang, H.A., 1972. Dynamic Model of a Cement kiln, Vol. 8, pp. 309 – 323,

Frish, V., Jeschar, R., 1983. Possibilities for Optimizing the Burning Process in Rotary Kiln, Vol. 10/83, pp.549-560,

Peray, K.E., 1986. The Rotary Cement Kiln, 2nd edition

Zhu, Y.C., 2001. Multivariable System Identification for process control, PERGAMON

Ljung, L., System Identification: Theory for the user. 1999 *Prentice- Hall,* 2nd edition

Nelles, O., 2001. Nonlinear system identification, Springer

Astrom, K. J. and Wittenmark, B., 1984. Computer Controlled Systems: Theory and Design, Prentice-Hall, Englewood Cliffs

Noshirvani, R., Identification of White Cement Rotary Cement Kiln, M.Sc Thesis, K.N.Toosi University Control Engineering Department of K.Ntoosi, 2005

Eykoff, P., 1974. System Identification: Parameter and state Estimation, John Wiley & Sons, New York

Shirvani, M., Dustary, M., Shahbaz, M. Eksiri, Z. Heuristic, 2004. Process Model Simplification inFrequency Response Domain, International of Engineering, Vol. 17/B, pp.31-52

ON THE LINEAR SCALE FRACTIONAL SYSTEMS
An Application of the Fractional Quantum Derivative

Manuel Duarte Ortigueira
UNINOVA and DEE of Faculdade de Ciências e Tecnologia da UNL
Campus da FCT da UNL, Quinta da Torre, 2825 – 114 Monte da Caparica, Portugal
mdortigueira@uninova.pt

Keywords: Linear Scale Invariant Systems, Fractional Linear Systems, Fractional Quantum Derivative.

Abstract: The Linear Scale Invariant Systems are introduced for both integer and fractional orders. They are defined by the generalized Euler-Cauchy differential equation. It is shown how to compute the impulse responses corresponding to the two regions of convergence of the transfer function. This is obtained by using the Mellin transform. The quantum fractional derivatives are used because they are suitable for dealing with this kind of systems.

1 INTRODUCTION

Braccini and Gambardella (1986) introduced the concept of “form-invariant” filters. These are systems such that a scaling of the input gives rise to a scaling of the output. This is important in detection and estimation of signals with unknown size requiring some type of pre-processing: for example edge sharpening in image processing or in radar signals. However in their attempt to define such systems, they did not give any formulation in terms of a differential equation. The Linear Scale Invariant Systems (LSIS) were really introduced by Yazici and Kashyap (1997) for analysis and modelling 1/f phenomena and in general the self-similar processes, namely the scale stationary processes. Their approach was based on an integer order Euler-Cauchy differential equation. However, they solved only a particular case corresponding to the all pole case. To insert a fractional behaviour, they proposed the concept of pseudo-impulse response. Here we avoid this procedure by presenting a fractional derivative based general formulation of the LSIS. We assume that the fractional LSIS is described by the general Euler-Cauchy differential equation

$$\sum_{i=0}^{N} a_i\, t^{\alpha_i}.y^{(\alpha_i)}(t) = \sum_{i=0}^{M} b_i\,.\, t^{\beta_i}.y^{(\beta_i)}(t) \qquad (1)$$

This equation is difficult to solve for any values for N or M and any derivative orders. However, when the derivative orders have the format

$$\alpha_i = \alpha+i \quad i=0, 1, 2, \ldots, N$$

and

$$\beta_i = \beta+i \quad i=0, 1, 2, \ldots, N$$

we obtain a simpler equation

$$\sum_{i=0}^{N} a_i\, t^{\alpha+i}.y^{(\alpha+i)}(t) = \sum_{i=0}^{M} b_i\,.\, t^{\beta+i}\, x^{(\beta+i)}(t) \qquad (2)$$

that we can solve with the help of the Mellin transform and using the fractional quantum derivative (Ortigueira, 2007, 2008). As we will show, the above equation allows us to obtain two transfer functions. Each of them has two terms that lead to two inverse functions. The impulse response is obtained by using the multiplicative convolution defined by (Bertran et al, 2000):

$$f(t)\vee g(t) = \int_0^\infty f(t/u)g(u)\frac{du}{u} \qquad (3)$$

Before going into the solution of equation (2), we are going to obtain the solution of the integer order equation corresponding to put $\alpha=\beta=0$ in (2). Then we will solve equation (2) for any α and β. This will be done in section 1. Other interesting results will be introduced in section 3. Finally we will present some conclusions.

2 THE EULER-CAUCHY EQUATION

2.1 The Integer Order Case

Consider a linear system represented by the differential equation

$$\sum_{i=0}^{N} a_i\, t^i.y^{(i)}(t) = \sum_{i=0}^{M} b_i\, .\, t^i\, x^{(i)}(t) \qquad (4)$$

where *x*(*t*) is the input, *y*(*t*) the output, and N and M are positive integers (M≤N). Usually a_N is chosen to be 1. We will assume that this equation is valid for every $t\in R^+$. Applying the Mellin transform to both sides of (3) we obtain (Gerardi, 1959;Bertran et al, 2000)

$$\sum_{i=0}^{N} a_i\, (-1)^i (s)_i\, Y(s) = \sum_{i=0}^{M} b_i\, .\, (-1)^i\, (s)_i\, X(s), \qquad (5)$$

from where we obtain a transfer function

$$H(s) = \frac{Y(s)}{X(s)} = \frac{\sum_{i=0}^{M} b_i\, (-1)^i\, (s)_i}{\sum_{i=0}^{N} a_i\, (-1)^i\, (s)_i} \qquad (6)$$

In this expression we need to transform both numerator and denominator into polynomials in the variable s. To do it we use the well known relation (Abramowitz and Stegun, 1972)

$$(x)_k = \sum_{i=0}^{k} (-1)^{k-i}\, s(k,i)\, x^i \qquad (7)$$

where s(,) represent the Stirling numbers of first kind that verify the recursion

$$s(n+1,m) = s(n,m-1) - ns(n,m) \qquad (8)$$

for $1 \le m \le n$ and with

$s(n,0) = \delta_n$ and $s(n,1) = (-1)^{n-1}(n-1)!$

With some manipulation, we obtain:

$$\sum_{i=0}^{N} a_i\, (-1)^i\, (x)_i = \sum_{i=0}^{N} \sum_{k=i}^{N} a_k\, (-1)^k s(k,i)\, x^i = \sum_{i=0}^{N} A_i\, x^i \qquad (9)$$

with the A_i coefficients given by

$$A_i = \sum_{i=k}^{N} a_k\, (-1)^k s(k,i) \qquad (10)$$

or in a matricial format

$$\mathbf{A} = \mathbf{S.a} \qquad (11)$$

where

$$\mathbf{A} = [A_0\ A_1\ \ldots\ \ldots\ A_N]^T \qquad (12)$$

$$\mathbf{S}=[\ s(i,j),\ i,j=0,1,\ \ldots,N] \qquad (13)$$

and

$$\mathbf{a} = [a_0\ a_1\ \ldots\ \ldots\ a_N]^T \qquad (14)$$

With this formulation, the transfer function is given by:

$$H(s) = \frac{\sum_{i=0}^{M} B_i\, s^i}{\sum_{i=0}^{N} A_i\, s^i} \qquad M\le N \qquad (15)$$

that is the quotient of two polynomials in s. In general H(s) has the following partial fraction decomposition

$$H(s) = \frac{B_M}{A_N} + \sum_{i=1}^{N} \sum_{j=1}^{m_i} \frac{a_{ij}}{(s-p_i)^j} \qquad (16)$$

The constant term only exists when M=N and its inversion gives a delta at t=1:

$$\mathcal{M}^{-1}\left[\frac{B_M}{A_N}\right] = \frac{B_M}{A_N}\delta(t-1) \qquad (17)$$

For inversion of a given partial fraction, we must fix the region of convergence Re(s) > Re(p_i) or Re(s) < Re(p_i) similar to identical situation found in the usual shift invariant systems with the Laplace transform. Let us assume that the poles are simple. Accordingly to each region of convergence we have (Bertran *et al*, 2000) respectively

$$\mathcal{M}^{-1}\left[\frac{1}{(s-p)}\right] = u(1-t).t^{-p} \qquad (18)$$

and

$$\mathcal{M}^{-1}\left[\frac{1}{(s-p)}\right] = u(t-1).t^{-p} \qquad (19)$$

By successive derivation in order to p we obtain the solution for higher order poles

$$\mathcal{M}^{-1}\left[\frac{1}{(s-p)^k}\right] = u(1-t).\frac{(-1)^{k-1}[\log(t)]^{k-1}}{(k-1)!}t^{-p} \qquad (20)$$

and

$$\mathcal{M}^{-1}\left[\frac{1}{(s-p)^k}\right] = u(t-1).\frac{(-1)^{k-1}[\log(t)]^{k-1}}{(k-1)!}t^{-p} \qquad (21)$$

We conclude that the response corresponding to an input δ(t-1) is given by:

$$h(t)=\frac{B_M}{A_N}\delta(t-1)+\sum_{i=1}^{N}\sum_{k=1}^{m_i} a_{ik}.\frac{(-1)^{k-1}[\log(t)]^{k-1}}{(k-1)!}t^{-p_i}\, w(t) \qquad (22)$$

where w(t) is equal to u(1-t) or to u(t-1), in agreement with the region of convergence adopted to invert (15). To compute the output to any function x(t) we only have to use the multiplicative convolution.

We must call the attention to the fact the point of application of the impulse is t=1 and not t=0, as it is the case of the shift-invariant systems.

2.2 The Fractional Quantum Derivative

To consider a more general case we must introduce the notion of fractional quantum derivative. This was not needed in the previous section because in the integer order case we only have one Mellin transform for $t^K f^{(K)}(t)$. This is not the situation in the fractional case. In fact we have two fractional derivatives given by {see appendix}:

$$D_q^{\alpha} f(t) = \lim_{q\to 1}\frac{\sum_{j=0}^{\infty}\begin{bmatrix}\alpha\\ j\end{bmatrix}_q (-1)^j q^{j(j+1)/2}\, q^{-j\alpha}\, f(q^j t)}{(1-q)^{\alpha}\, t^{\alpha}} \qquad (23)$$

and

$$D_{q^{-1}}^{\alpha} f(t) = \lim_{q\to 1}\frac{\sum_{j=0}^{\infty}\begin{bmatrix}\alpha\\ j\end{bmatrix}_q (-1)^j q^{j(j-1)/2}\, f(q^{-j} t)}{(1-q^{-1})^{\alpha}\, t^{\alpha}} \qquad (24)$$

These derivatives have the same Mellin transform in the integer order case, but in the general their Mellin transforms are given by:

$$\mathcal{M}\left[D_q^{\alpha} f(t)\right] = \frac{\Gamma(1-s+\alpha)}{\Gamma(1-s)}\, F(s-\alpha) \qquad (25)$$

valid for Re(s) < min(0,α)+1, in the first case and by

$$\mathcal{M}\left[D_{q^{-1}}^{\alpha} f(t)\right] = (-1)^{\alpha}.\frac{\Gamma(s)}{\Gamma(s-\alpha)}\, F(s-\alpha) \qquad (26)$$

valid for Re(s) > max(0,α), in the second case. It is interesting that the first corresponds to the anti-causal case when working in the Laplace transform context, while the second corresponds to the causal one.

2.3 The Fractional Order Equation

Consider now a linear system represented by the fractional differential equation

$$\sum_{i=0}^{N} a_i\, t^{\alpha+i}.y^{(\alpha+i)}(t) = \sum_{i=0}^{M} b_i\, .\, t^{\beta+i}\, x^{(\beta+i)}(t) \qquad (27)$$

where α and β are real numbers. With the Mellin transform we obtain two different transfer functions depending on the derivative we use, (23) or (24). From (23) we have:

$$H(s) = \frac{\sum_{i=0}^{M} b_i\,(-1)^i\,(s+\beta)_i}{\sum_{i=0}^{N} a_i\,(-1)^i\,(s+\alpha)_i}\cdot\frac{\Gamma(1-s-\alpha)}{\Gamma(1-s)}\,\frac{\Gamma(1-s)}{\Gamma(1-s-\beta)} \qquad (28)$$

Proceeding as in 2.1 we have

$$H(s) = \frac{\sum_{i=0}^{M} B_i\,(s+\beta)^i}{\sum_{i=0}^{N} A_i\,(s+\alpha)^i}\cdot\frac{\Gamma(1-s-\alpha)}{\Gamma(1-s-\beta)} \qquad (29)$$

So, the transfer function in (29) has two parts; the first is similar to (25) aside a translation on the pole and zero positions. Its inverse has the format:

$$h(t)=\frac{B_M}{A_N}\delta(t-1)+t^{\alpha}\sum_{i=1}^{N}\sum_{k=1}^{m_i} C_{ik}.\frac{(-1)^{k-1}[\log(t)]^{k-1}}{(k-1)!}t^{-p_i}\, u(t-1) \qquad (30)$$

where the p_i, i=1,2, …, N are the poles. We must remark that it does not depend explicitly on β. The second factor in (29) leads to a new convolutional factor needed to compute the complete solution of (27). So, we have to compute the inverse Mellin transform of

$$H_a(s) = .\frac{\Gamma(1\text{-}s\text{-}\alpha)}{\Gamma(1\text{-}s\text{-}\beta)} \qquad (31)$$

To do it we can always choose an integration path on the left of all the poles. Computing this integral, we obtain:

$$h_a(t) = \frac{1}{\Gamma(\alpha\text{-}\beta)} t^{\beta}(t - 1)^{\alpha\text{-}\beta\text{-}1}u(t\text{-}1) \qquad (32)$$

So, the impulse response corresponding to (29) is the convolution of (30) and (32). By simplicity, assume that all the poles are simple. In this case, the impulse response is given by:

$$\begin{aligned} h(t) = {} & \frac{B_M}{A_N} \frac{1}{\Gamma(\alpha\text{-}\beta)} t^{\beta}(t - 1)^{\alpha\text{-}\beta\text{-}1}u(t\text{-}1) + \\ & + t^{\alpha} \sum_{i=1}^{N} C_i . \frac{\Gamma(1\text{-}p_i)}{\Gamma(\,\alpha\text{-}\beta\text{-}p_i\text{+}1)} t^{\text{-}p_i}\, u(t\text{-}1) \end{aligned} \qquad (33)$$

Choosing the other region of convergence we have

$$H(s) = \frac{\sum_{i=0}^{M} B_i\, (s\text{+}\beta)^i}{\sum_{i=0}^{N} A_i\, (s\text{+}\alpha)^i} .(\text{-}1)^{\beta\text{-}\alpha} \frac{\Gamma(s\text{+}\beta)}{\Gamma(\, s\text{+}\alpha)} \qquad (34)$$

The first factor has as inverse the expression:

$$\begin{aligned} h(t) = {} & \frac{B_M}{A_N} \delta(t\text{-}1) + \\ & + t^{\alpha} \sum_{i=1}^{N} \sum_{k=1}^{m_i} C_{ik} . \frac{(\text{-}1)^k[\log(t)]^{k\text{-}1}}{(k\text{-}1)!} t^{\text{-}p_i}\, u(1\text{-}t) \end{aligned} \qquad (35)$$

For the second we proceed as before. Now the integration path is in the right half complex plane. We obtain

$$h_a(t) = - \frac{1}{\Gamma(\alpha\text{-}\beta)} t^{\beta}(t - 1)^{\alpha\text{-}\beta\text{-}1}u(1\text{-}t) \qquad (36)$$

To compute the final impulse response we only have to convolve the two expressions as we did in the other case. We obtain, for the simple pole case

$$\begin{aligned} h(t) = {} & - \frac{B_M}{A_N} \frac{1}{\Gamma(\alpha\text{-}\beta)} t^{\beta}(t - 1)^{\alpha\text{-}\beta\text{-}1}u(1\text{-}t) \text{ -} \\ & t^{\alpha} \sum_{i=1}^{N} C_i . \frac{\Gamma(\beta\text{-}\alpha\text{+}p_i)}{\Gamma\,(p_i)} t^{\text{-}p_i} u(1\text{-}t) \end{aligned} \qquad (37)$$

It is interesting to verify that (33) and (37) behavior like the usual anti-causal and causal systems. When $\text{Re}(p_i) < 0$, (30) increases without bound while (35) decreases. If $\text{Re}(p_i) > 0$, we verify the reverse situation. This means that we can use the well known Routh-Hurwitz test to study the stability of LSIS.

2.4 Particular Cases

2.4.1 $\alpha = \beta$

If $\alpha=\beta$, the second terms in (29) and (34) is equal to 1, implying that the complete impulse response is given by (30) and (35).

2.4.2 $\alpha = 0$ and $\beta \neq 0$

This case is very interesting since it is similar to the situation treated by Yazici and Kashyap. With $\alpha=0$, (30) and (35) do not depend explicitly on β and they are similar to the integer order case. The dependence on β appears only in the second therm.

2.4.3 $\alpha \neq 0$ and $\beta = 0$

This situation is more involved, since both terms of the impulse response depend on α. We can obtain the general impulse response by putting $\beta=0$ into (30), (32), (35), and (36).

3 THE EIGENFUNCTIONS AND FREQUENCY RESPONSE

Consider relation (3) and assume that one of the functions is the impulse response of the system (1) and the other is a power function $t^{\text{-}\sigma}$, $\sigma\in C$. It is not hard to show that

$$h(t) \vee t^{\text{-}\sigma} = H(\sigma).t^{\text{-}\sigma} \qquad (38)$$

Leading us to conclude that the power function is the eigenfunction of the LSIS. In particular we can write:

$$h(t) \vee t^{\text{-}j\nu} = H(\nu).t^{\text{-}j\nu} \qquad (39)$$

and $H(\nu)$ will be the frequency response of the system, considering that our "cisoids" have the format

$$c(t) = e^{\text{-}j\nu\log(t)} \qquad (40)$$

that verify:

$$c(t) = c(at) \qquad (41)$$

provided that

$$a = e^{2\pi/\nu} \qquad (42)$$

defining the scale periodicity. These results show that the output of a LSIS to a cisoid is a cisoid. For a cosine signal, as input, the output y(t) is given by

$$y(t) = |H(\nu)|.\cos[2\pi\nu\log(t)+\varphi(\nu)] \qquad (43)$$

where $\varphi(\nu)$ is the phase spectrum of the system.

4 CONCLUSIONS

In this paper, we introduced the general formulation of the linear scale invariant systems through the fractional Euler-Cauchy equation. To solve this equation we used the fractional quantum derivative concept and the help of the Mellin transform. As in the linear time invariant systems we obtained two solutions corresponding to the use of two different regions of convergence. We presented other interesting features of the LSIS, namely the frequency response. We made also a brief study of the stability.

ACKNOWLEDGEMENTS

This work was supported by the Portuguese Foundation for Science and Technology through the program FEDER/POSC.

REFERENCES

Abramowitz, M. and Stegun, I. (1972) Stirling Numbers of the First Kind., §24.1.3 in Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, 9th printing. New York: Dover, p. 824.

Andrews, G.E., Askey, R., and Roy, R. (1999). Special functions. Cambridge University Press.V.

Ash, J.M., Catoiu, S., and Rios-Collantes-de-Terán, R., (2002), "On the nth Quantum Derivative," J. London Math. Soc. (2) 66, 114-130.

Bertran, J. Bertran, P., and Ovarlez, J.P., (2000), "The Mellin Transform", in "The Transforms and Applications Handbook", 2nd ed,, Editor-in-Chief A.D. Poularikas, CRC Press,.

Braccini, C. and Gambardella, G., (1986), "Form-invariant linear filtering: Theory and applications," IEEE Trans. Acoust., Speech, Signal Processing, vol.ASSP-34, no. 6, pp. 1612–1628,.

Gerardi, F.R., (1959) "Application of Mellin and Hankel transforms to networks with time varying parameters," IRE Trans. Circuit Theory, vol. CT-6, pp. 197–208.

Kac, V. and Cheung, P. (2002) "Quantum Calculus", Springer,.

Koornwinder, T.H., (1999) "Some simple applications and variants of the q-binomial formula," Informal note, Universiteit van Amsterdam,.

Ortigueira, M. D., (2006), A coherent approach to non integer order derivatives, Signal Processing Special Section: Fractional Calculus Applications in Signals and Systems, vol. 10, pp. 2505-2515.

Ortigueira, M.D., (2007) "A Non Integer Order Quantum Derivative", Symposium on Applied Fractional Calculus (SAFC07), Industrial Engineering School (University of Extremadura), Badajoz (Spain), October 15-17.

Ortigueira, M.D., (2008) "A fractional Quantum Derivative", proceedings of the IFAC Fractional Differentiation and its Applications conference, Ankara, Turkey, 05 - 07 November.

Yazici, B. and Kashyap, R.L., (1997) "A Class of Second-Order Stationary Self-Similar Processes for 1/f Phenomena," IEEE Transactions on Signal Processing, vol. 45, no. 2.

APPENDIX - QUANTUM DERIVATIVE FORMULATIONS

- **Incremental Ratio Formulation**

The normal way of introducing the notion of derivative is by means of the limit of an incremental ratio that in the forward case reads

$$D_hf(t) = \lim_{h\to 0}\frac{f(t) - f(t-h)}{h} \qquad (a.1)$$

By repeated application, this definition leads to the derivative of any integer order that can be generalized to any real or complex order by the well known forward Grünwald-Letnikov fractional derivative (Ortigueira, 2006):

$$D_h^{\alpha}f(z) = \lim_{h\to 0+}\frac{\sum_{k=0}^{\infty}(-1)^k\binom{\alpha}{k}f(z-kh)}{h^{\alpha}} \qquad (a.2)$$

An alternative derivative valid only for t>0 or t<0 is the so-called quantum derivative (Kac and Cheug, 2002). Let Δ_q be the following incremental ratio:

$$\Delta_qf(t) = \frac{f(t) - f(qt)}{(1-q)t} \qquad (a.3)$$

where q is a positive real number less than 1 and f(t) is assumed to be a causal type signal. The corresponding derivative is obtained by computing the limit as q goes to 1

$$D_qf(t) = \lim_{q\to 1}\frac{f(t) - f(qt)}{(1-q)t} \qquad (a.4)$$

This derivative uses values of the variable below t. We can introduce another one that uses values above t. It is defined by

$$D_{q^{-1}} f(t) = \lim_{q\to 1} \frac{f(q^{-1}t) - f(t)}{(q^{-1} - 1)t} \qquad \text{(a.5)}$$

The repeated application of (a.3) followed by the limit computation leads to the N^{th} order derivative (Ash et al, 2002;Koornwinder, 1999):

$$D_q^N f(t) = \lim_{q\to 1} \frac{\sum_{j=0}^{N} \begin{bmatrix} N \\ j \end{bmatrix}_q (-1)^j q^{j(j+1)/2} q^{-jN} f(q^j t)}{(1-q)^N t^N} \qquad \text{(a.6)}$$

where we introduced the q-binomial coefficients

$$\begin{bmatrix} \alpha \\ i \end{bmatrix}_q = \frac{[\alpha]_q!}{[i]_q![\alpha-i]_q!} \qquad \text{(a.7)}$$

with $[\alpha]_q$ given by

$$[\alpha]_q = \frac{1-q^\alpha}{1-q} \qquad \text{(a.8)}$$

Using the q-binomial theorem (Kac and Cheug, 2002), the Mellin transform, and the Pochhamer symbol we conclude that:

$$\mathcal{M}\left[\lim_{q\to 1} \frac{\sum_{j=0}^{N} \begin{bmatrix} N \\ j \end{bmatrix}_q (-1)^j q^{j(j+1)/2} q^{-jN} f(q^j t)}{(1-q)^N t^N}\right] \qquad \text{(a.9)}$$

$$= (1-s)_N F(s-N)$$

$$= \frac{\Gamma(1-s+N)}{\Gamma(1-s)} F(s\text{-}N)$$

The previous results are readily generalised for the case of a real order, α, (Ortigueira,2007; Ortigueira,2008) leading to a Grunwald-Letnikov like fractional quantum derivative:

$$D_q^\alpha f(t) = \lim_{q\to 1} \frac{\sum_{j=0}^{\infty} \begin{bmatrix} \alpha \\ j \end{bmatrix}_q (-1)^j q^{j(j+1)/2} q^{-j\alpha} f(q^j t)}{(1-q)^\alpha t^\alpha} \qquad \text{(a.10)}$$

that is similar to the one proposed by Salam (1966).

In (a.10) the fractional q-binomial coefficients are given by

$$\begin{bmatrix} \alpha \\ j \end{bmatrix}_q = \frac{[1-q^\alpha]_q^j}{[j]_q} \qquad \text{(a.11)}$$

The Mellin transform of (a.10) reads

$$\mathcal{M}\,[D_q^\alpha f(t)\,] = \frac{\Gamma(1-s+\alpha)}{\Gamma(1-s)} F(s\text{-}\alpha) \qquad \text{(a.12)}$$

valid for Re(s) < min(0,α)+1. This relation allows us to obtain an integral representation of the fractional quantum derivative, as we will see later. As referred before, in (a.10) we are using values of the variable less than t. In the following we will consider the other case. The repeated application of (a.5) leads to the N^{th} order derivative:

$$D_{q^{-1}}^N f(t) = \lim_{q\to 1} \frac{\sum_{j=0}^{N} \begin{bmatrix} N \\ j \end{bmatrix}_q (-1)^j q^{j(j-1)/2} f(q^{-j} t)}{(1-q^{-1})^N t^N} \qquad \text{(a.13)}$$

The Mellin transform gives:

$$\mathcal{M}\,[D_{q^{-1}}^N f(t)\,] = (1\text{-}s)_N F(s\text{-}N) \qquad \text{(a.14)}$$

that coincides with (a.9) as expected. To generalize the above results for any order, we substitute α for N in the above expressions. We have from (a.10):

$$D_{q^{-1}}^\alpha f(t) = \lim_{q\to 1} \frac{\sum_{j=0}^{\infty} \begin{bmatrix} \alpha \\ j \end{bmatrix}_q (-1)^j q^{j(j-1)/2} f(q^{-j} t)}{(1-q^{-1})^\alpha t^\alpha} \qquad \text{(a.15)}$$

and finally

$$\mathcal{M}\,[D_{q^{-1}}^\alpha f(t)\,] = (-1)^\alpha \cdot \frac{\Gamma(s)}{\Gamma(s-\alpha)} F(s\text{-}\alpha) \qquad \text{(a.16)}$$

valid for Re(s) > max(0,α). Remark the difference relatively to (a.12) mainly in the region of convergence.

- **Integral Formulations**

The two Mellin transforms in (a.12) and (a.16) lead to different integral representation of fractional derivatives by computing the corresponding inverse functions.

The inverse $h_b(t)$ of $\frac{\Gamma(s)}{\Gamma(s-\alpha)}$ is obtained from (Andrews *et al*,1999):

$$\frac{\Gamma(s)\Gamma(-\alpha)}{\Gamma(s-\alpha)} = \int_0^1 \tau^{s-1}\,(1-\tau)^{-\alpha-1}d\tau \qquad (a.17)$$

Provided that Re(s)>0 and Re(α)<0. This leads immediately to

$$h_b(t) = \frac{(-1)^{\alpha}}{\Gamma(-\alpha)}(1 - t)^{-\alpha-1}u(1-t) \qquad (a.18)$$

u(t) is the Heaviside unit step. A similar procedure to obtain the inverse $h_a(t)$ of $\frac{\Gamma(1-s+\alpha)}{\Gamma(1-s)}$ gives

$$\frac{\Gamma(1-s+\alpha)\Gamma(-\alpha)}{\Gamma(1-s)} = \int_0^1 \tau^{1-s+\alpha}\,(1-\tau)^{-\alpha-1}d\tau \qquad (a.19)$$

With a variable change inside the integral, we obtain:

$$h_a(t) = \frac{1}{\Gamma(-\alpha)}(t - 1)^{-\alpha-1}u(t-1) \qquad (a.20)$$

To compute in integral formulations of the derivatives corresponding to (a.12) and (a.16) we remark that the inverse Mellin transform of F(s-α) is given by:

$$\mathcal{M}^{-1}[F(s-\alpha)] = t^{-\alpha}f(t) \qquad (a.21)$$

and use the convolution (3). With (a.12) and (a.16) we obtain the following integral formulations, valid for Re(α) < 0.

$$D_b^{\alpha}f(t) = -\frac{t^{-\alpha}}{\Gamma(-\alpha)}\cdot\int_0^1 f(t/\tau)\,(1-\tau^{-1})^{-\alpha-1}d\tau \qquad (a.22)$$

and

$$D_a^{\alpha}f(t) = \frac{t^{-\alpha}}{\Gamma(-\alpha)}\int_1^{\infty} f(t/\tau)\,(\tau^{-1}-1)^{-\alpha-1}\,d\tau \qquad (a.23)$$

signals. Although we obtained these results for α<0, they remain valid for other values of α, since $\frac{\Gamma(s)}{\Gamma(s-\alpha)}$ and $\frac{\Gamma(1-s+\alpha)}{\Gamma(1-s)}$ are analytic in the regions of convergence and we can fix an integration path independent of α. This can be confirmed by expanding (a.22) and (a.23) and transforming each term of the series.

POSTERS

MICRO-ENERGY PULSE POWER SUPPLY WITH NANOSECOND PULSE WIDTH FOR EDM

Fang Ji[1], Yong-bin Zhang[2], Guang-min Liu[3] and Jian-guo He[3]
Institute of Machinery Manufacturing Technology, Mianshan Road 64#, Mianyang, Sichuan, 621900, China
[1]fred110@sohu.com, [2]wang1970ok@sohu.com, [3]endeavorup@163.com

Keywords: Electrical discharging machining, Micro-energy, Pulse power supply.

Abstract: Micro-energy pulse power supply is required in order to manufacture workpiece with micro-nano meter precision in electrical discharging machining (EDM). The paper analyzes two kinds of typical pulse power supplies and their important elements. Afterwards, another one kind of new micro-energy pulse power supply is presented. The experiments and analysis have been done for the new power supply. Accordingly, some important circuits have been modified, for example, the discharging circuit and driving circuit for the metallic oxide semiconductor field effect transistor (MOSFET). The modification improves the performance of the new pulse power supply so that the pulse width of the new pulse power supply could be less than that of the typical pulse power supply for electrical discharging machining. The least pulse width is obtained. It is less than 60 nanoseconds and its least energy of single pulse is less than 10-6 joule. Subsequently, the pulse waveform is adjusted considering the impedance matching of the discharging circuit in order that the pulse waveform has no oscillation and no overshot. The adjusted pulse waveform is good to detect discharging status correctly and sensitively.

1 INTRODUCTION

The independent pulse power supply and RC pulse power supply are two typical pulse power supplies applied in electrical discharging machining. The independent pulse power supply has the advantages: discharging frequency may be high, pulse parameters may be adjustable, self-adapted control may be easy. But it still has the disadvantages: maintaining voltage limits the energy of single pulse to decrease, the energy of single pulse is more than 10^{-7} joule. The disadvantages make independent pulse power supply difficult to manufacture workpiece with micro-nano meter precision. The other typical pulse power supply is RC pulse power supply. It is easy to obtain small energy of 10^{-7} joule of single pulse. But, it is difficult to adjust the pulse parameters. It has no channel to release residual charge between two electrodes. It is not easy to control energy of pulse. Electrical arc discharging happens frequently and the discharging consistency is not good. The researches show that the machining mass in micro-nano meter scale needs a kind of micro-energy pulse power supply. Its pulse parameters must be easy to control and its lowest energy of single pulse must reach 10^{-7} joule.

The independent pulse power supply is shown in Figure.1. The energy of single pulse W_0 is related to instantaneous discharging voltage $u(t)$, instantaneous discharging current $i(t)$ and pulse width T. Their relation may be written as $W_0 = \int_0^T u(t)i(t)dt$. According to the relation, it is obvious that there are three ways to decrease energy of single pulse: decreasing the voltage, decreasing the current or increasing the frequency. However, there exists maintaining voltage which is the least voltage to discharge between electrodes. The discharging voltage must be larger than the maintaining voltage. It limits the decrease of the discharging voltage. Additionally, the increase of frequency may be restricted by the frequency response of the MOSFET. Therefore, it is difficult to acquire less energy of single pulse for independent pulse power supply and the lowest pulse energy can only reach 10^{-6} joule. The RC pulse power supply is shown in Figure.2. The energy W_{RC} stored in the capacitor may be described as the relation between the capacitance C of capacitor, the capacitance C' of circuit and the discharging voltage U :

$W_{RC} = \frac{1}{2}(C + C')U^2$. Hereby, there are two ways to decrease the energy of single pulse: decreasing the capacitance and the voltage. Normally speaking, the first way seems better. But, the capacitance C' usually varies from 100pF to 10000pF in the circuit and it is hard to reduce. Thus, decreasing discharging voltage may be the most important way to reduce the energy of single pulse. The new research shows that the discharging voltage will not be limited by maintaining voltage and may be as low as 7 volts for RC pulse power supply. The minimum energy of single pulse may reach 10^{-7} joule. Whereas, the present RC pulse power supply is difficult to control and residual charge is easy to accumulate between electrodes, which is not good for consistency of machining. The low discharging voltage will brings the discharging distance to be close which is not good to remove the leftover.

At present, there are many researchers who are developing the micro-energy pulse power supply for EDM. The least pulse width is 90 nanoseconds developed by Zhao, but there exists obvious electromagnetic oscillation; The least pulse width is 80 nanoseconds developed by Han, but the width is width of current, not width of voltage. It is well-known that the pulse width of current is less than that of voltage because of the discharging delay. The paper presents one kind of micro-energy pulse power supply which integrates the advantages of both independent pulse power supply and RC pulse power supply. Its least energy of single pulse can reach 10^{-7} joule. It has a special circuit to release residual charge between electrodes.

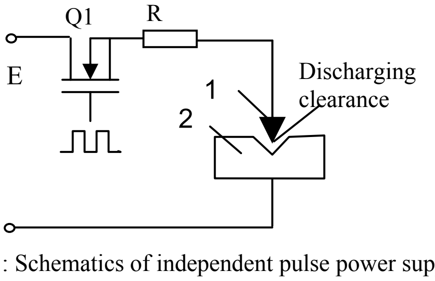


Figure 1: Schematics of independent pulse power supply.

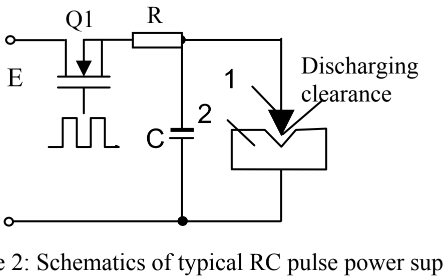


Figure 2: Schematics of typical RC pulse power supply.

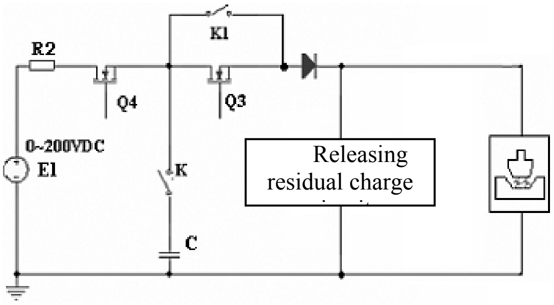


Figure 3: Shematics of micro-energy pulse power supply with nanosecond pulse width.

2 PRINCIPLE OF THE MICRO-ENERGY PULSE POWER SUPPLY

The micro-energy pulse power supply presented in the paper is shown in Figure.3. The current will charge capacitor C through MOSFET Q4 and resistor R2 when the switch K1 is disconnected and the switch K is connected. The energy in capacitor is decided by the charging time and it can influence the machining mass of single pulse. Then, the energy in capacitor will transfer to the discharging clearance between workpiece and tool electrode when Q4 is disconnected and the Q3 is connected. Afterwards, Q3 will be disconnected and the releasing residual charge circuit will remove the residual energy in order to avoid unnecessary discharging between electrodes. At last, the releasing residual charge circuit will be disconnected. The whole work period of single pulse is over and the next period may begin. A programmable logic element is applied in the system circuit to control the MOSFETs. Thus, some logic operations are done by hardware rapidly, which may reduce the delay time and decrease the pulse width. In addition, a special high-speed micro-control unit (MCU) is configured as counter for pulse so that fuzzy control may be done according to the number of pulse. The main elements in the system circuit are high-speed MOSFETs. They can work at high frequency. They influence the minimum energy of single pulse and the machining efficiency of the pulse power supply. However, there exists nonlinearity between gate voltage and source voltage during charging and discharging because of capacitance in MOSFET. Therefore, the internal wastage will increase and the reliability will decrease. This is a disadvantage. But, it can be reduced by high-speed driving circuit for the MOSFET. The driving circuit has instantaneous strong current so that the MOSFET may be connected or disconnected quickly. Thus, the pulse width less than 60 nanoseconds may be obtained.

3 EXPERIMENTS AND ANALYSIS

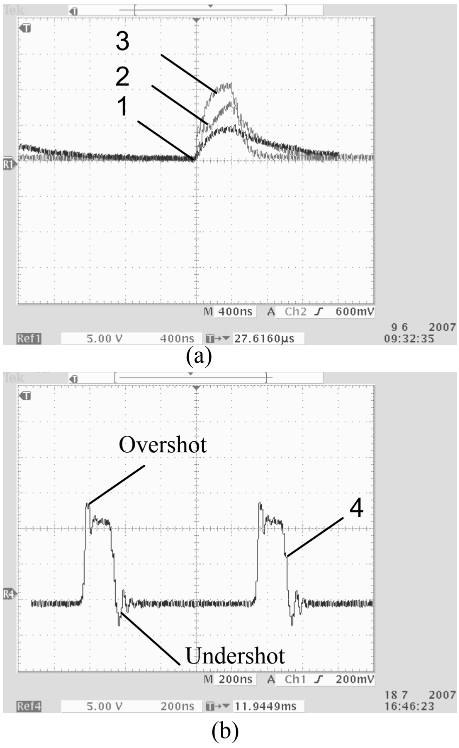
Some experiments have been done in the paper. Then, analysis and improvements have been given for optimized pulse power supply, for example, the modification of driving circuit, the comparison of MOSFETs, the impedance matching, discharging circuit and detection of discharging status, etc.

3.1 Modification of the Driving Circuit

The system circuit adopts a driving chip with complementary emitter follower which has strong capability to drive gate of MOSFET. Otherwise, the impedance of the driving circuit is low. Thus, the charging and discharging for the gate of MOSFET can be finished quickly. There are four waveforms in Figure.4. They are different because their impedances are not equal. The impedances vary from large to small by the order 1 to 4. It is obvious that decreasing impedance may improve the waveform. The fourth waveform is adopted in the paper. The overshot at rising edge and undershot at falling edge are good for connecting and disconnecting of MOSFET.

3.2 Comparison of the MOSFET

Comparison experiments of MOSFETs have been done to select the best kind of MOSFET when the parameters of the system circuit are of no difference. Three kinds of MOSFETs are selected for further experiments after some previous experiments. The corresponding waveforms of different MOSFETs are shown in Figure.5. It can be found that there still exist some undershots for the 1st and 2nd waveform. The 3rd waveform is the best one. Hereby, the 3rd MOSFET is fixed for the pulse power supply.



(a)

(b)

Figure 4: Different waveforms of driving circuit with different impedances. The impedance varies from large to small by the order 1 to 4.

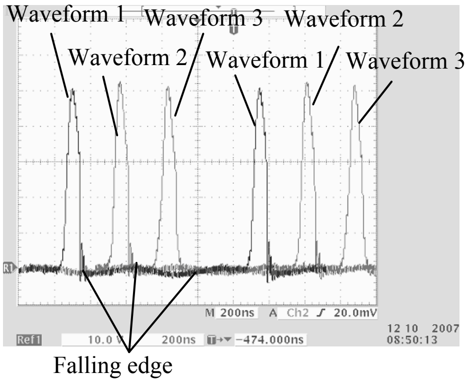
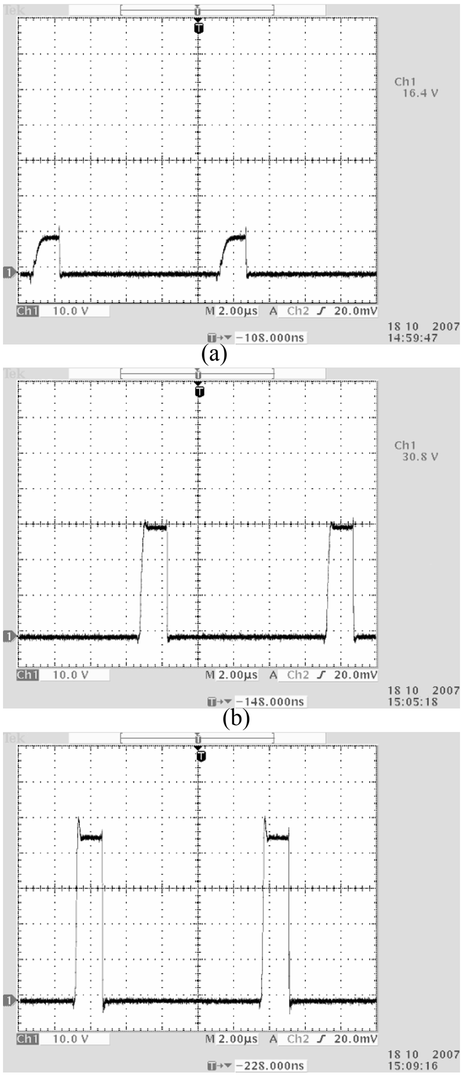


Figure 5: Different waveforms of different MOSFETs.

3.3 Matching of the Impedance

There are two special results during the experiments of pulse power supply: Firstly, the waveform of MOSFET will be different when the voltage is different, even though the electronical elements are all same. It is shown in Figure.6. The overshot will appear at the rising edge with the increase of voltage. The reason is that the rise of voltage results in the rise of varying rate of current. The capacitor will be charged more quickly, which makes overshot at rising edge of voltage. Secondly, the waveform of MOSFET will also be different when electronical element is changed for different machining currents. It is shown in Figure.7. The overshot will appear for some elements. The reason is that the impedances are different for different elements. The less the impedance is, the larger the overshot will be.

However, their falling edges are similar because of the releasing residual charge circuit. Therefore, the impedance must be matched for the optimized waveform and detection accuracy of discharging status.



(a)

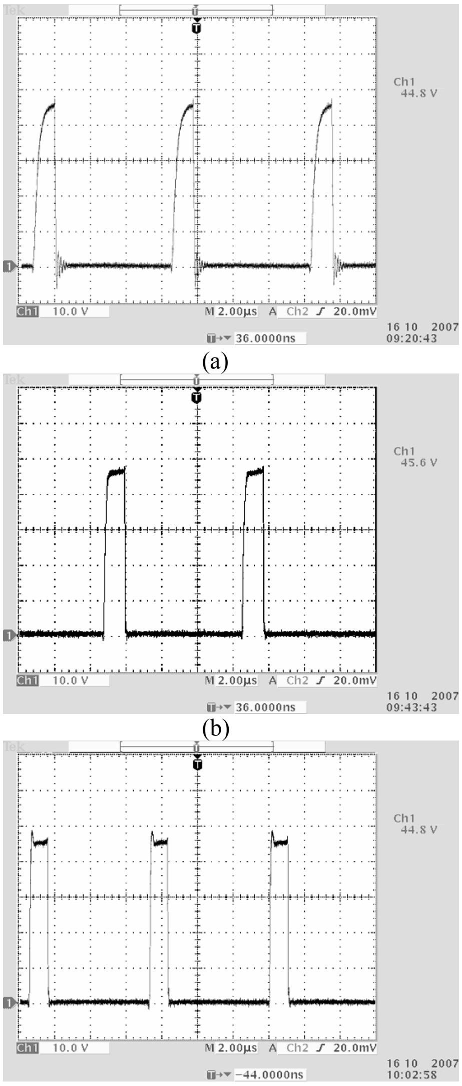
(b)

(c)

Figure 6: Waveforms with different voltages and same electronical element.

3.4 Discharging Experiment at High Frequency

The least pulse width of the micro-energy pulse power supply described in the paper can be less than 60 nanoseconds after some experiments and optimization above. The discharging experiments have been done subsequently. The open waveform and discharging waveform are shown in Figure.8. It is easy to watch the spark between the anode and the cathode during the experiments. There exist some discharging marks on the surface of the workpiece. The energy of single pulse can be calculated by the equation $W_0 = \int_0^T u(t)i(t)dt$ and reaches to 10^{-7} joule.

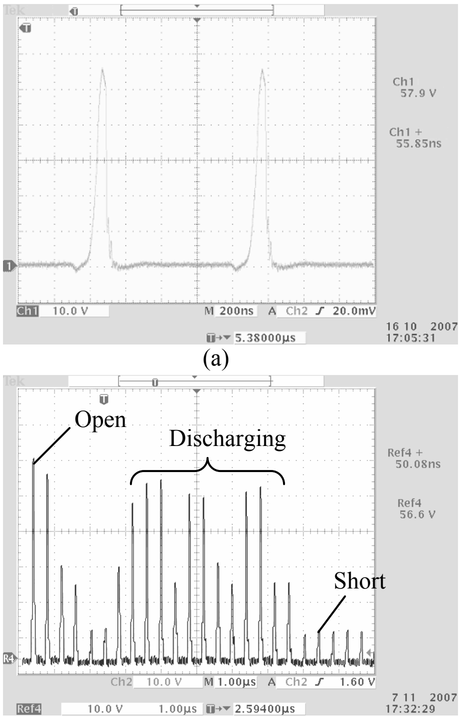


(a)

(b)

(c)

Figure 7: Waveforms with same voltage and different electronical elements.



(a)

(b)

Figure 8: Open waveforms and discharging waveforms of the pulse power supply.

3.5 Detection of Discharging Status

Detection of discharging status is very important in the pulse power supply. Its result will be fed back to the control center. It provides the main information used to adjust the parameters of pulse power supply timely. The corresponding circuit must be modified accurately. But, the signal from discharging clearance is periodic and changed quickly, which will bring variance, even oscillation, to the detection circuit. The signal is shown in Figure.9. Thus, some filters are applied in the circuit. Some noise is limited and the signal is improved evidently. It is shown in Figure.10.

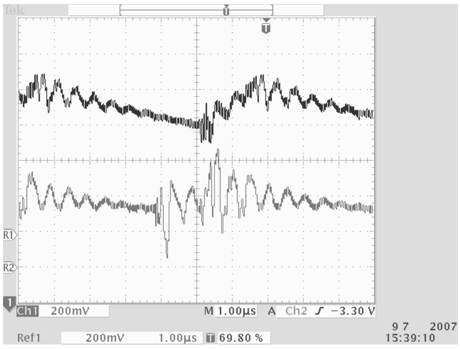


Figure 9: Voltage of open clearance with interference.

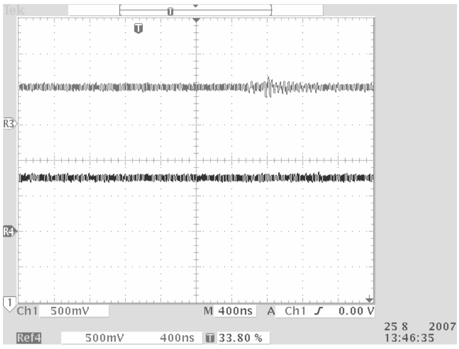


Figure 10: Voltage of open clearance without interference.

4 CONCLUSIONS

The micro-energy pulse power supply is the key part of EDM in micro-nano meter scale. The paper analyzes the characters of two present typical pulse power supplies of EDM, gives the driving circuit consisted of low impedance element and complementary emitter follower. The paper also adds active releasing residual charge circuit, selects the best MOSFET and matches the impedance for different waveforms. At last, optimized waveform is obtained without overshot and undershot. The minimum energy of the pulse power supply reaches 10^{-7} joule and the least pulse width is less than 60 nanoseconds.

REFERENCES

Ao Ming-wu, Zhang Yong, Li Zhi-yong, 2003, Research on a micro-energy pulse power source used by micro-EDM, *Aviation Precision Manufacturing Technology*.

Chi Guan-xin, Di Shi-chun, 2004, Study on key technology for intricate micro parts on MWEDM, *Manufacturing Technology and Machine Tool*.

Y.S. Wong a,., M. Rahmana, H.S. Lima, H. Hanb, N. Ravib, 2003, Investigation of micro-EDM material removal characteristics, *Journal of Materials Processing Technology*.

He Guang-ming, Zhao Wan-sheng, 1999, Research of Nanosecond-class pulse generator, *Electromachining*.

Han Fu-zhu, Chen li, Zhou Xiao-guang, 2005, Foundational research on pulse generator technology in micro-EDM, *Electromachining & Mould*.

Cao Feng-guo s, 2005, *Electro discharge machining technology*, Beijing: Chemical Industry Pres.

Pei Jing-yu, Guo Chang-ning, Deng Qi-lin, 2004, Dual-Channel MOSFET Sub- Microsecond Micro- Energy Pulse Power Source Used in Electrical Discharge Machining, *Journal of Shanghai Jiaotong University*.

A NEW DECONVOLUTION METHOD BASED ON MAXIMUM ENTROPY AND QUASI-MOMENT TRUNCATION TECHNIQUE

Monika Pinchas
Department of Electrical and Electronic Engineering, Ariel University Center of Samaria, Ariel 40700, Israel
monika.pinchas@gmail.com

Ben Zion Bobrovsky
Department of Electrical Engineering-Systems, Tel-Aviv University, Tel Aviv 69978, Israel

Keywords: Blind equalization, Blind deconvolution, Non-linear adaptive filtering.

Abstract: In this paper we present a new blind equalization method based on the quasi-moment truncation technique and on the Maximum Entropy blind equalization method presented previously in the literature. In our new proposed method, fewer moments of the source signal are needed to be known compared with the previously presented technique. Simulation results show that our new proposed algorithm has better equalization performance compared with Godard's and Lazaro's *et al.* algorithm.

1 INTRODUCTION

We consider a blind equalization problem in which we observe the output of an unknown, possibly non-minimum phase, linear system from which we want to recover its input using an adjustable linear filter (equalizer). The problem of blind equalization arises comprehensively in various applications such as digital communications, seismic signal processing, speech modeling and synthesis, ultrasonic non-destructive evaluation, and image restoration (Feng and Chi, 1999). Recently, a new blind equalization algorithm was proposed (Pinchas and Bobrovsky, 2006) with improved equalization performance compared with (Godard, 1980) and (Lazaro et al., 2005). It is valid for the real and complex (where the real and imaginary parts are independent) valued case. This new blind equalization method (Pinchas and Bobrovsky, 2006) is based on the Maximum Entropy technique and on some known moments of the source signal. The problem arises when these moments or part of them are unknown. In that case the blind equalization method (Pinchas and Bobrovsky, 2006) can not be used. Obviously, when using approximated moments instead of the real ones, the equalization performance might get worse and in some cases even lead to unacceptable performance. The quasi-moment truncation technique is related to the Hermite polynomials where the high-order central moments are approximated in terms of lower order central moments (Bover, 1978). Although the quasi-moment truncation technique (Bover, 1978) is well known in the non-linear optimal filtering theory (Bover, 1978), it is not yet been used in the field of blind equalization combined with the Maximum Entropy technique. In this paper we present a new blind equalization method based on the quasi-moment truncation technique and on the Maximum Entropy blind equalization method (Pinchas and Bobrovsky, 2006). Fewer moments of the source signal are needed to be known compared with (Pinchas and Bobrovsky, 2006). Simulation results will show that our new proposed algorithm has better equalization performance compared with Godard's (Godard, 1980) and Lazaro's *et al.* (Lazaro et al., 2005) algorithm. The paper is organized as follows: After having described the system under consideration in Section II, we describe in Section III the quasi-moment truncation technique which we use in this paper for approximating the unknown source moments. In Section IV we present our simulation results and Section V is our conclusion.

2 SYSTEM DESCRIPTION

The system under consideration is illustrated in Fig.1, where we make the following assumptions:
1. The input sequence $x(n)$ consists of zero mean

real or complex (where the real and imaginary part of $x(n)$ are independent) random variables with an unknown even symmetric probability distribution.
2. The unknown channel $h(n)$ is a possibly nonminimum phase linear time-invariant filter in which the transfer function has no “deep zeros”, namely, the zeros lie sufficiently far from the unit circle.
3. The equalizer $c(n)$ is a tap-delay line.
4. The noise $w(n)$ is an additive Gaussian white noise.
5. The function $T[\cdot]$ is a memoryless nonlinear function. It satisfies the analyticity condition:
$T(z_1 + jz_2) = T_1(z_1) + jT_2(z_2)$ where z_1, z_2, are the real and imaginary part of the equalized output respectively.

The transmitted sequence $x(n)$ is transmitted through the channel $h(n)$ and is corrupted with noise $w(n)$. Therefore, the equalizer's input sequence $y(n)$ may be written as:

$$y(n) = x(n) * h(n) + w(n) \tag{1}$$

where ”$*$” denotes the convolution operation. This sequence (1) is then equalized with an equalizer $c(n)$. The equalizer's output sequence $z(n)$ may be written as:

$$\begin{gathered} z(n) = x(n) * h(n) * c(n) + w(n) * c(n) = \\ x(n) + p(n) + \tilde{w}(n) \end{gathered} \tag{2}$$

where $p(n)$ is the convolutional noise and $\tilde{w}(n) = w(n) * c(n)$. In this paper, we consider the equalizer proposed by (Pinchas and Bobrovsky, 2006) where the equalizer's taps are updated according to:

$$\begin{gathered} c_l(n+1) = c_l(n) - \mu W y^*(n-l) \qquad \text{with} \\ W = \left[(W_1 + W_2) - z[n]\right] \\ W_1 = E\left[x_1/z_1\right]\left[\frac{\left(z_1[n] E\left[x_1/z_1\right]\right)}{\left\langle (z_1)^2 \right\rangle_n}\right] \\ W_2 = jE\left[x_2/z_2\right]\left[\frac{\left(z_2[n] E\left[x_2/z_2\right]\right)}{\left\langle (z_2)^2 \right\rangle_n}\right] \\ \left\langle z_s^2 \right\rangle_n = (1-\beta)\left\langle z_s^2 \right\rangle_{n-1} + \beta \cdot (z_s)_n^2 \end{gathered} \tag{3}$$

where $()^*$ is the conjugate of $()$, μ is a positive step-size parameter, l stands for the l-th tap of the equalizer, $\langle\rangle$ stands for the estimated expectation, $\left\langle z_s^2 \right\rangle_0 > 0$ $(s = 1,2)$, β is a positive stepsize parameter and $E[x_s/z_s]$ $(s = 1,2)$ is the conditional expectation derived in (Pinchas and Bobrovsky, 2006) with the use of the Maximum Entropy density approximation technique. This blind equalization algorithm (3) depends on some known moments of the source signal through the expression of the conditional expectation given in (Pinchas and Bobrovsky, 2006). The problem arises when we do not know these moments or we know only a part of them. In that case we can not use the algorithm. In the following we will show how we solve this problem and still obtain satisfying equalization performance compared with (Godard, 1980) and (Lazaro et al., 2005).

3 MOMENT APPROXIMATION

In this section we use the quasi-moment truncation technique (Bover, 1978) for approximating the unknown source moments. In the following we consider the real valued case. The quasi-moment truncation technique is related to the Hermite polynomials where the high-order central moments are approximated in terms of lower order central moments (Bover, 1978). According to (Bover, 1978), one way of achieving this is by expressing the probability density function $f_x(x)$ as an infinite series expansion in which the coefficients are known in terms of central moments. Then truncation approximations is done by assuming that high-order coefficients in this expansion are negligible. This would seem likely to occur when the basis for the expansion is an appropriate set of orthogonal polynomials (Bover, 1978). A natural choice of expansion basis is the Hermite polynomials (Bover, 1978) which was used by Kuznetsov, Stratonovich and Tikhonov (Kuznetsov et al., 1960) who introduced the name “quasi-moment” for the expansion coefficients. Thus following (Bover, 1978), the probability density function $f_x(x)$ is expressed as:

$$f_x(x) = \frac{1}{\sqrt{2\pi}\sigma_x} \exp\left(-\frac{x^2}{2\sigma_x^2}\right) \sum_{L=0}^{\infty} \frac{b_L}{L!} H_L(x) \tag{4}$$

where b_L are the quasi-moments and $H_L(x)$ are the Hermite polynomials defined by:

$$H_L(x) = \exp\left(\frac{x^2}{2\sigma_x^2}\right)\left(-\frac{d}{dx}\right)^L \left[\exp\left(-\frac{x^2}{2\sigma_x^2}\right)\right] \tag{5}$$

According to (Bover, 1978), we may deduce quite simple expressions for the quasi-moments in terms of central moments by using the property, proved by (Appel and Feriet, 1926), that the Hermite polynomials are orthogonal with their adjoint polynomials, with respect to a Gaussian weight function. By a straight forward manipulation we may find that any quasi-moment is equal to the expectation of the corresponding adjoint Hermite polynomial (Bover, 1978), namely:

$$\begin{gathered} b_L = < G_L(x) > \qquad \text{where} \\ G_L(x) = \exp\left(\frac{\widetilde{x}^2\sigma_x^2}{2}\right)\left(-\frac{d}{d\widetilde{x}}\right)^L \exp\left(-\frac{\widetilde{x}^2\sigma_x^2}{2}\right) \\ \text{with} \qquad \widetilde{x} = \frac{x}{\sigma_x^2} \end{gathered} \tag{6}$$

In the following is a list of the first six one-dimensional quasi-moments calculated by (Bover, 1978):

$$\begin{gathered} b_0 = 1; \quad b_1 = 0; \quad b_2 = 0; \quad b_3 = \langle x^3 \rangle \\ b_4 = \langle x^4 \rangle - 3\langle x^2 \rangle^2; \quad b_5 = \langle x^5 \rangle - 10\langle x^2 \rangle \langle x^3 \rangle \\ b_6 = \langle x^6 \rangle - 15\langle x^2 \rangle \langle x^4 \rangle + 30\langle x^2 \rangle^3 \end{gathered} \tag{7}$$

Now, assuming for instance that b_6 is negligible ($b_6 = 0$), an approximation for the six-th central moment in terms of lower order central moments is obtained.

4 SIMULATION

In this section we investigate the equalization performance by simulation where we use the residual ISI (intersymbol interference) as a measure of performance. Note that the ISI is often used as a measure of performance in equalizers' applications. In the following, we denote "MaxEnt" as the algorithm described by (3) with the Lagrange multipliers given in (Pinchas and Bobrovsky, 2006) where the required source moments are known. The step-size parameters for this method were denoted as μ and β and we substituted $E[z_s^2] = E[x_s^2]$ for initialization. The equalizer taps for Godard's algorithm (Godard, 1980) were updated according to:

$$\begin{gathered} c_l(n+1) = c_l(n) - \\ \mu_G \left(|z(n)|^2 - \frac{E\left[|x(n)|^4\right]}{E\left[|x(n)|^2\right]} \right) z(n) y^*(n-l) \end{gathered} \tag{8}$$

where μ_G is the step-size. The equalizer taps for algorithm (Shalvi and Weinstein, 1990) were updated according to:

$$\begin{gathered} c'_i(n+1) = c''_i(n) + \mu_{SW} \cdot \operatorname{sgn}\Upsilon(x) |z(n)|^2 z(n) \cdot \\ y^*(n-i) \quad \text{where} \quad c''_i(n) = \left(1 \Big/ \sqrt{\sum_i |c'_i|^2} \right) c'_i \end{gathered} \tag{9}$$

where $c''_i(n)$ is the vector of taps after iteration, $c''_i(0)$ is some reasonable initial guess, μ_{SW} is the step-size and $\Upsilon(x) = E\left[|x|^4\right] - 2E^2\left[|x|^2\right] - \left|E\left[x^2\right]\right|^2$ is the kurtosis associated to x. In the following, we denote algorithm (Shalvi and Weinstein, 1990) as SW. The equalizer taps for algorithm (Lazaro et al., 2005) were updated according to:

$$\begin{gathered} c_l(n+1) = c_l(n) - \\ \mu_{par} \left(\left(\frac{1}{N_{sym}} \right) \left(\sum_{k=1}^{N_{sym}} \tilde{K}'_\sigma \left(|z(n)|^2 - F(\sigma) |x_k|^2 \right) \right) \right) \cdot \\ z(n) y^*(n-l) \end{gathered} \tag{10}$$

where μ_{par} is the step-size, $\tilde{K}'_\sigma(z)$ is the derivative of $\tilde{K}_\sigma(z)$ which is the Parzen window kernel of size σ and $F(\sigma)$ is the compensation factor that depends on the kernel size. In (Lazaro et al., 2005) the Gaussian kernel with standard deviation σ was used for $\tilde{K}_\sigma(z)$: $\tilde{K}_\sigma(z) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{z^2}{2\sigma^2}\right)$. In the following, we denote algorithm (Lazaro et al., 2005) as SQD. We denote "MaxEntA" as the algorithm described by (3) with the Lagrange multipliers given in (Pinchas and Bobrovsky, 2006) where some of the required source moments are approximated according to the quasi-moment truncation technique (7). The step-size parameters for this method were denoted as μ_A and β_A and we substituted $E[z_s^2] = E[x_s^2]$ for initialization unless otherwise stated. We used in our simulation a **16QAM source** (a modulation using ± {1,3} levels for in-phase and quadrature components). Two channels were considered. **Channel1** (initial ISI = 0.44): The channel parameters were determined according to (Shalvi and Weinstein, 1990):

$$h_n = \{0 \quad \text{for} \quad n < 0; \quad -0.4 \quad \text{for} \quad n = 0 \quad 0.84 \cdot 0.4^{n-1} \quad \text{for} \quad n > 0\} \tag{11}$$

Channel2 (initial ISI = 1.402): The channel parameters were taken according to (Lazaro et al., 2005): $h_n = (0.2258, 0.5161, 0.6452, 0.5161)$.

For Channel1 a 13-tap equalizer was used. For Channel2 we used an equalizer with 21 taps. In our simulation, the equalizers were initialized by setting the center tap equal to one and all others to zero. The step-size parameters μ, μ_A, μ_G, β, β_A, μ_{SW}, μ_{par} were chosen for fast convergence with low steady state ISI. For the 16QAM source input propagating through Channel2, the performance of Godard's and SQD algorithm were reproduced following (Lazaro et al., 2005). For the 16QAM modulation source, two Lagrange multipliers (λ_2, λ_4) were used by the "MaxEnt" and "MaxEntA" algorithm. For the "MaxEntA" algorithm, m_6 was approximated according to the quasi-moment truncation method while the other moments m_4 and m_2 were assumed to be known. Figure 2 shows the equalization performance of "MaxEnt" and "MaxEntA" compared with (Lazaro et al., 2005) and (Shalvi and Weinstein, 1990) for the 16QAM source constellation propagating through channel1. The performance is expressed in terms of residual ISI as a function of iteration number. Figure 3 shows the equalization performance of "MaxEntA" with and without the use of initial samples for the initialization phase compared with (Lazaro et al., 2005) and (Godard, 1980) for the 16QAM source constellation propagating through channel2. According to the simulated results, our new proposed algorithm "MaxEntA" has improved equalization performance compared with (Godard, 1980), (Shalvi and Weinstein, 1990) and (Lazaro et al., 2005).

5 CONCLUSIONS

We have derived in this paper a new blind equalization method based on the quasi-moment truncation technique and on the Maximum Entropy blind equalization method (Pinchas and Bobrovsky, 2006). In our proposed algorithm, fewer moments of the source signal are needed to be known compared with (Pinchas and Bobrovsky, 2006). Simulation results indicate that the new proposed algorithm has improved equalization performance compared with (Godard, 1980) and (Lazaro et al., 2005).

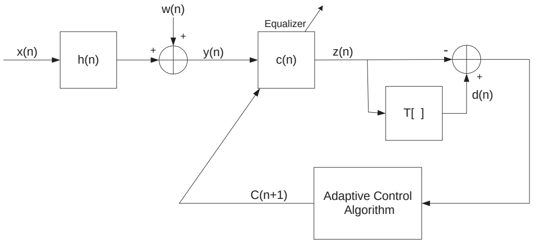


Figure 1: Baseband communication system.

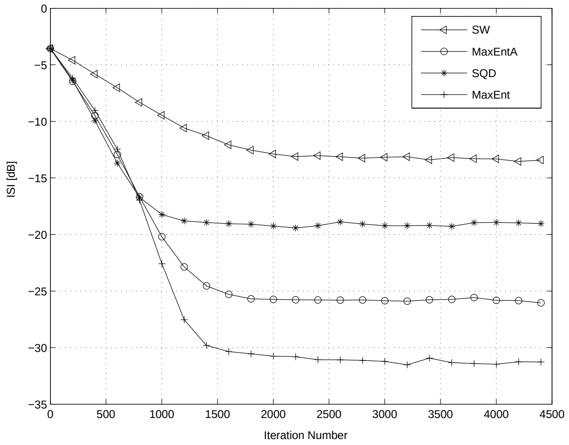


Figure 2: Performance comparison between equalization algorithms for a 16QAM source input going through channel1. The averaged results were obtained in 100 Monte Carlo trials for a SNR of 30 (dB). The step-size parameters were set to: $\mu_{SW} = 2.5e$-5, $\mu = 3e$-4, $\beta = 2e$-4, $\mu_A = 3.5e$-4, $\beta_A = 4e$-4 and $\mu_{par} = 2.5e$-4. In addition we set $F(\sigma) = 1$, $\sigma = 15$ and ε to 0.5, 0 for MaxEnt and MaxEntA respectively.

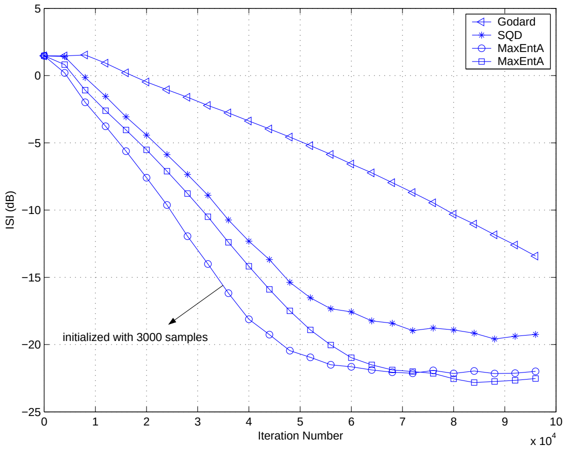


Figure 3: Performance comparison between equalization algorithms for a 16QAM source input going through channel2. The averaged results were obtained in 50 Monte Carlo trials for a SNR of 30 (dB). The step-size parameters were set to: $\mu_A = 2e$-4, $\beta_A = 2e$-6, $\mu_A = 2.5e$-4 for “o” , $\beta_A = 2e$-6 for “o”, $\mu_{par} = 1e$-4 and $\mu_G = 1e$-5. In addition we set $\varepsilon = 0.5$, $F(\sigma) = 1$ and $\sigma = 15$.

REFERENCES

Appel, P. and Feriet, J. K. D. (1926). Fonctions hypergeometriques et hyperspheriques. In *polynomes d'Hermite 367-369*. Gauthier-Villars, Paris.

Bover, D. C. C. (1978). Moment equation methods for nonlinear stochastic systems. In *Journal of Mathematical Analysis and Applications 65 306-320*.

Feng, C. and Chi, C. (1999). Performance of cumulant based inverse filter for blind deconvolution. In *IEEE Transaction on Signal Processing 47 (7) 1922-1935*. IEEE.

Godard, D. (1980). Self recovering equalization and carrier tracking in two-dimentional data communication system. In *IEEE Transaction Communication 28 (11) 1867-1875*. IEEE.

Kuznetsov, P. I., Stratonovich, R. L., and Tikhonov, V. I. (1960). Quasi-moment functions in the theory of random process. In *Theor. Probability Appl. 5 80-97*.

Lazaro, M., Santamaria, I., Erdogmus, D., Hild, K. E., Pantaleon, C., and Principe, J. C. (2005). Stochastic blind equalization based on pdf fitting using parzen estimator. In *IEEE Transaction on Signal Processing 53 (2) 696-704*. IEEE.

Pinchas, M. and Bobrovsky, B. Z. (2006). A maximum entropy approach for blind deconvolution. In *Signal Processing Vol. 86, issue 10 2913-2931*. Elsevier.

Shalvi, O. and Weinstein, E. (1990). New criteria for blind deconvolution of nonminimum phase systems (channels). In *IEEE Trans. Information Theory 36 (2), 312-321*.

ABOUT THE DECOMPOSITION OF RATIONAL SERIES IN NONCOMMUTATIVE VARIABLES INTO SIMPLE SERIES

Mikhail V. Foursov
IRISA/Université de Rennes–1, Campus Universitaire de Beaulieu, 35042 Rennes Cedex, France
mikhail.foursov@irisa.fr

Christiane Hespel
IRISA/INSA de Rennes, 20, avenue des Buttes de Coësmes, 35043 Rennes Cedex, France
christiane.hespel@insa-rennes.fr

Keywords: Formal power series, Generating series, Rational series, Joint matrix block–diagonalization, Dynamical systems.

Abstract: Similarly to the partial fraction decomposition of rational fractions, we provide an approach to the decomposition of rational series in noncommutative variables into simpler series. This decomposition consists in splitting the representation of the rational series into simpler representations. Finally, the problem appears as a joint block–diagonalization of several matrices. We present then an application of this decomposition to the integration of dynamical systems.

1 INTRODUCTION

This article deals with the problem of splitting a rational formal power series into simple series. We present first well–known results on decomposition of rational series in a single variable and on reduced linear representations of a rational series in noncommutative variables.

Fliess showed that decomposition of rational formal power series can be done by joint block-diagonalization of several matrices. This is a difficult problem which was approached by numerous researchers such as Gantmacher, Jordan, Dunford and Jacobi.

The decomposition into simple series has many different applications in the dynamical system theory (such as subsystem independence, integration or stability) and in the automata theory, among others. We illustrate the application to the integration of dynamical systems.

2 PRELIMINARIES

In this paper, we consider a rational series s with coefficients in the field $K = \mathbb{C}$. In some sections, K can be taken as a semi–ring or as a commutative field.

2.1 Decomposition of Rational Series in a Single Variable into Simple Series

A rational series s in a single variable can be rewritten as a rational fraction (Gantmacher, 1966).

Theorem 2.1. *Let $s = \sum_{j=0}^{\infty} s_j X^{j+1} \in K[[X]]$ be a formal power series with coefficients in a field K of characteristic 0. Then there are 2 polynomials $P, Q \in K[X]$, such that*

$$deg(Q) < deg(P), \quad \frac{Q}{P} = \sum_{j=0}^{\infty} \frac{s_j}{X^{j+1}} \qquad (1)$$

if and only if there is an integer $p \in \mathbb{N}$ such that the ranks of the Hankel matrices of orders k, $\forall k \geq p$, are all equal to p.

In this case there exist polynomials P of degree p and Q of degree at most $p-1$. The minimal possible degree of P is p, and the pair (P, Q) is completely determined by these degree conditions and the condition that P is monic. The polynomials P and Q are then prime.

The proof of this theorem is based on the resolution of a system of linear equations obtained by identifying the coefficients of X^l. Let us remark that the finiteness condition on the rank of the Hankel matrix of s expresses the recognizability of s, that is the rationality, for a single variable. This rational fraction can be easily split up into simple fractions of the form

$s_i = \frac{a_i}{(1-\alpha_i X)^{r_i}}$ where $a_i, \alpha_i \in \mathbb{C}, r_i \in \mathbb{N}$, s_i being expanded as a rational simple series.

Remark. A rational series can be considered as a weighted automaton (also known as automaton with multiplicity). The previous decomposition of s as $s = \sum_{i \in I} s_i$ appears as a decomposition of the weighted automaton $\mathrm{A_s}$ of dimension r into $\cup_{i \in I} \mathrm{A_{s_i}}$, where $\mathrm{A_{s_i}}$ are simple independent automata of dimension r_i such that

$$\begin{cases} dim(\mathrm{A_{s_i}}) = & r_i \\ \sum_{i \in I} r_i = & r \end{cases} \tag{2}$$

2.2 Reduced Linear Representation of Rational Series in Noncommutative Variables

2.2.1 Series in Noncommutative Variables

These definitions and notations are from (Berstel and Reutenauer, 1988; Reutenauer, 1980; Salomaa and Soittola, 1978; Schützenberger, 1961). K is a semi–ring.

Definition 2.1. *(Formal power series in noncommutative variables)*

1. *An alphabet X is a nonempty finite set. Elements of X are letters. The free monoid X^* generated by the alphabet X is the set of finite words $X_{i_1} \cdots X_{i_l}$, where $X_{i_j} \in X$, including the empty word denoted by 1. The set X^* is a monoid with respect to concatenation.*
2. *A formal power series s in noncommutative variables is a function*
$$s: \ X^* \to K \tag{3}$$
The coefficient $s(w)$ of the word w in the series s is denoted by $\langle s|w \rangle$.
3. *The set of formal power series s over X with coefficients in K is denoted by $K\langle\langle X \rangle\rangle$. A structure of semi–ring is defined on $K\langle\langle X \rangle\rangle$ by the sum and the Cauchy product. Two external operations (left and right products) from K to $K\langle\langle X \rangle\rangle$ are also defined. The set of polynomials is denoted by $K\langle X \rangle$.*

2.2.2 Rational Series in Noncommutative Variables

Definition 2.2. *(Rational formal power series in noncommutative variables)*

1. *The rational operations in $K\langle\langle X \rangle\rangle$ are the sum, the product, two external products as well as the Kleene star operation defined by $T^* = \sum_{n \geq 0} T^n$ for a proper series T (i.e. such that $\langle T|1 \rangle = 0$).*
2. *A subset of $K\langle\langle X \rangle\rangle$ is rationally closed if it is closed under the rational operations. The smallest rationally–closed subset containing a subset $E \subseteq K\langle\langle X \rangle\rangle$ is called the rational closure of E.*
3. *A series s is rational if s is an element of the rational closure of $K\langle X \rangle$.*

2.2.3 Recognizable Series in Noncommutative Variables

We propose several equivalent definitions (Berstel and Reutenauer, 1988; Fliess, 1977; Fliess, 1974; Fliess, 1976; Jacob, 1980), K being a commutative field.

Definition 2.3. *(Recognizable formal power series in noncommutative variables)*

1. *A series $s \in K\langle\langle X \rangle\rangle$ is recognizable if there exists an integer $N \geq 1$, a monoid morphism*
$$\mu : X^* \to K^{N*N} \tag{4}$$
and 2 matrices $\lambda \in K^{1*N}$ and $\gamma \in K^{N*1}$ such that
$$\forall w \in X^*, \ \langle s|w \rangle = \lambda\mu(w)\gamma. \tag{5}$$
2. *A series $s \in K\langle\langle X \rangle\rangle$ is recognizable if there exists an integer N, the rank of its Hankel matrix $H(s) = (\langle s|w_1.w_2 \rangle)_{w_1,w_2 \in X^*}$. The first row of $H(s)$ indexed by the word 1 describes s. The other rows are the remainders of s by a word w. For instance, the row L_{X_1} represents the right remainder of s by X_1, denoted by $s \triangleright X_1$.*
3. *A series $s \in K\langle\langle X \rangle\rangle$ is recognizable if it is described by a finite weighted automaton obtained from its Hankel matrix remainders.*

Definition 2.4. *The triple (λ, μ, γ) is called a linear representation of s. The representation with minimal dimension is called the reduced linear representation.*

2.2.4 Theorem of Schützenberger

For a series in several noncommutative variables, the theorem of Schützenberger proves the equivalence between the notions of rationality and of recognizability (Schützenberger, 1961; Berstel and Reutenauer, 1988).

Theorem 2.2. *A formal series is recognizable if and only if it is rational.*

2.2.5 Finite Weighted Automaton Obtained from a Rational Series

This method is developed in (Hespel, 1998). It is based on the following theorem (Fliess, 1976; Jacob, 1980).

Theorem 2.3. *A formal series $s \in \mathbb{R}\langle\langle X\rangle\rangle$ is recognizable if and only if its rank N is finite. Then it is recognized by a $\mathbb{R}$–matrix automaton $M = (N, \gamma, \lambda, \mu)$. Two sets of words $\{g_i\}_{1\le i\le N}$ and $\{d_j\}_{1\le j\le N}$, whose lengths are $< N$, can be determined so that the application χ from X^* to $\mathbb{R}^{N\times N}$ defined by*

$$(\chi(w))_{i,j} = \langle s|g_i.w.d_j\rangle \qquad (6)$$

satisfies $\chi(w) = \chi(1)\mu(w)$ with $\chi(1)$ invertible.

1. The method consists in extracting from the Hankel matrix $H(s)$ (whose rank is N) a system B of N row vectors $(L_{w_i})_{i\in I}$ (resp. N column vectors $(C_{w_j})_{j\in J}$), indexed by some words of minimum length, such that their determinant is nonzero and such that every row (resp. every column) of $H(s)$ can be expressed as a linear combination of elements of B. These relations allow us to define $\forall X_k \in X$ the matrices $\mu(X_k)$ describing the action of the letter X_k on the row vector L_{w_i} (resp. the column vector C_{w_j}). The first row (resp. the first column) of B defines λ. γ is the initial vector $(1\,0\cdots 0)^T$. The series s can thus be written
$$s = \sum_{w\in X^*}\langle s|w\rangle = \sum_{w\in X^*}\lambda\mu(w)\gamma \qquad (7)$$
2. We define, based on the basis B and matrices $\mu(X_i)$, γ and λ, a finite weighted (left or right) automaton $A = \{X, Q, I, A, \tau\}$ such that
   - X is the alphabet,
   - the state set is $Q = \{L_{w_i}\}_{i\in I}$ representing $\{s \triangleright w_i\}_{i\in I}$ (resp. $Q = \{C_{w_j}\}_{j\in J}$ representing $\{w_j \triangleleft s\}_{j\in J}$),
   - the first row (resp. the first column) I of B is the initial state,
   - every transition between states belonging to τ is labeled by a letter $X_i \in X$ and labeled by the coefficient appearing in the linear dependence relation,
   - A is the final state set; it is the set of rows L_w (resp. the columns C_w) of B such that $\langle s|w\rangle \ne 0$.

3 DECOMPOSITION OF RATIONAL SERIES : PRINCIPLE

3.1 Theoretical Results

In his thesis (Fliess, 1977), M.Fliess gives the idea of a unique decomposition of the reduced matrix representation μ associated to a rational series s into the direct sum of a finite number of simple representations. His idea is based on the Krull–Schmidt theorem.

Let us recall some definitions and notations (Berstel and Reutenauer, 1988; Fliess, 1977).

Let $s \in K\langle\langle X\rangle\rangle$ be a rational series. Let us denote by $\{N, \lambda, \mu(X^*), \gamma\}$, or rather by μ, its reduced matrix representation. The coefficients of s satisfy

$$\langle s|w\rangle = \lambda\mu(w)\gamma, \quad \forall w \in X^* \qquad (8)$$

For a decomposition of μ

$$\mu = \oplus_{i=1}^{k}\mu_i \qquad (9)$$

the associated decompositions of the vectors λ and γ are

$$\lambda = \oplus_{i=1}^{k}\lambda_i, \qquad \gamma = \oplus_{i=1}^{k}\gamma_i \qquad (10)$$

The series s is then split up into $s = \sum_{i=1}^{k} s_i$, where every rational series satisfies

$$s_i = \sum_{w\in X^*}\big(\lambda_i\mu_i(w)\gamma_i\big)w \qquad (11)$$

Among $\{s_i\}_{1\le i\le k}$ there can exist a subfamily with indices $J \subseteq \{1,\cdots,k\}$ such that $\forall j \in J$, the representation μ_j is nilpotent.

- *A representation μ_i is nilpotent* if and only if $\forall w \in X^+,\ \mu_j(w)$ is nilpotent.

Using Levitzki theorem (Kaplanski, 1969), the semi–group of nilpotent matrices $\{\oplus_{j\in J}\ \mu_j(w),\ w \in X^+\}$ is simultaneously triangulable. Particularly, for every word w of sufficient length, $\oplus_{j\in J}\ \mu_j(w)$ is the zero matrix. Then the sum $\sum_{j\in J}\ s_j$ of the series associated to this decomposition into nilpotent matrices is a polynomial representing the polynomial part of s.

Let us consider now the representations which cannot be decomposed and which are not nilpotent.

- Such a representation μ_i is associated with a simple series s_i.
- Two series s_1 and s_2 are called relatively prime if and only if
$$\begin{gathered}\forall\alpha,\ \beta \in C\backslash\{0\},\\ rank(\alpha s_1 + \beta s_2) = rank(s_1) + rank(s_2)\end{gathered} \qquad (12)$$

We can express the following theorem (Fliess, 1977)

Theorem 3.1. *K being a field, there is a unique way for decomposing every rational series $s \in K\langle\langle X\rangle\rangle$ into the sum of its polynomial part and of some simple rational relatively prime series.*

3.2 Approaches of the Simultaneous Decomposition of Matrices $\{A_i\}_{i\in I}$

We restrict the number of matrices to two in order to simplify the explanations. The problem is the following : to provide a simultaneous decomposition of A_1 and A_2 into a nilpotent part A_{1_n}, A_{2_n} and a block–diagonalizable part A_{1_d}, A_{2_d}, in some basis.

This problem is difficult. We present some approaches from Gantmacher, Jordan, Dunford and Jacobi.

1. **First Approach : Gantmacher**

   Gantmacher considers the linear pencil $A_1 + \lambda A_2$ of the matrices A_1, A_2. By using elementary transformations, ((Gantmacher, 1966), tome 1, Chapter 2), the original regular/singular pencil can be reduced to a quasi–diagonal canonical form ((Gantmacher, 1966), tome 2, Chapter 12). The original pencil $A_1 + \lambda A_2$ and the canonical pencil $A'_1 + \lambda A'_2$ are then equivalent but generally not similar : there exist some regular matrices P, Q such that $A'_1 + \lambda A'_2 = P(A_1 + \lambda A_2)Q$ but generally $Q \neq P^{-1}$.

2. **Second Approach : Jordan, Dunford**

   These methods are suitable for a single matrix. The Jordan's method consists in computing 2 regular matrices P, Q and irreducible block diagonal matrices A'_1, A'_2 such that

   $$A_1 = P^{-1}A'_1P, \ A_2 = Q^{-1}A'_2Q. \qquad (13)$$

   So one can use the Jordan decomposition A'_1 and A'_2 of each matrix in order to initialize a simultaneous decomposition in block diagonal matrices of suitable size. The knowledge of the eigenspaces (E_{1_i}) and (E_{2_i}) of A_1 and A_2 allows to set some bounds on the size of the blocks.

   The Dunford decomposition into a diagonalizable part and a nilpotent part can be provided from the Jordan decomposition.

3. **Approach by Jacobi Algorithms**

   When the sizes of the decomposition blocks are known, the method consists in providing a joint block–diagonalizer. This matrix is iteratively computed as a product of Givens rotations. The convergence of this algorithm is proven but not necessary to obtain an optimal solution.

4 DECOMPOSITION OF RATIONAL SERIES IN PRACTICE

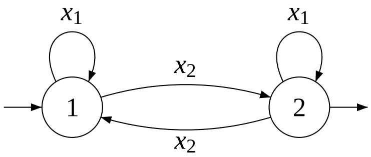
Theorem 4.1. *A rational series can be decomposed into a sum of simpler series using matrix joint block–decomposition.*

Proof. Let s be a rational series $s = \sum_{w\in X^*} \langle s|w\rangle = \sum_{w\in X^*} \lambda\mu(w)\gamma$. For a simultaneous change of basis matrix P for $\mu(x_{i_j})_{i_j}$, we have

$$\begin{aligned}\langle s|x_{i_1}\cdots x_{i_l}\rangle = \lambda\mu(x_{i_1})\cdots\mu(x_{i_l})\gamma = \\ = \lambda P\mu'(x_{i_1})P^{-1}\cdots P\mu'(x_{i_l})P^{-1}\gamma \\ = (\lambda P)\mu'(x_{i_1})\cdots\mu'(x_{i_l})(P^{-1}\gamma) = \\ = \lambda_P\mu_P(x_{i_1})\cdots\mu_P(x_{i_l})\gamma_P\end{aligned} \qquad (14)$$

Thus, when $\mu'(x_{i_1}), \cdots, \mu'(x_{i_l})$ are decomposed into block–diagonal matrices, we obtain the decomposition of s into corresponding simpler series. □

Example 1. A representation of the series is given by the finite weighted automaton



The actions of the letters x_1 and x_2 are given by the matrices

$$\mu(x_1) = \begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix} \quad \text{and} \quad \mu(x_2) = \begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix} \qquad (15)$$

The initial vector is

$$\gamma = \begin{pmatrix}1\\ 0\end{pmatrix} \qquad (16)$$

and the covector is

$$\lambda = \begin{pmatrix}0 & 1\end{pmatrix}. \qquad (17)$$

The eigenvalues of $\mu(x_2)$ are $\lambda_1 = 1$ and $\lambda_2 = -1$. In the basis B of the eigenvectors, the matrices $\mu(x_1)$ and $\mu(x_2)$ are

$$\mu(x_1)_P = \begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix} \text{ and } \mu(x_2)_P = \begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix} \qquad (18)$$

The initial vector is now

$$\gamma_P = \begin{pmatrix}1/2\\ 1/2\end{pmatrix} \qquad (19)$$

and the covector is

$$\lambda_P = \begin{pmatrix}1 & -1\end{pmatrix}. \qquad (20)$$

Thus this series can be decomposed into series s_1 and $s_2 : s = s_1 + s_2$. The representation of s_1 is

$$\mu_1(x_1) = (1),\ \mu_1(x_2) = (1),\ \gamma_1 = (1/2),\ \lambda_1 = (1). \tag{21}$$

For s_2 we have

$$\mu_2(x_1) = (1),\ \mu_2(x_2) = (-1),\ \gamma_1 = (1/2),\ \lambda_1 = (-1). \tag{22}$$

Example 2. Now let us consider the series with the following representation. The actions of the letters x_1 and x_2 are given by the matrices

$$\mu(x_1) = \begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix} \quad \text{and} \quad \mu(x_2) = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix} \tag{23}$$

The initial vector is

$$\gamma = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \tag{24}$$

and the covector is

$$\lambda = \begin{pmatrix} 0 & 1 \end{pmatrix} \tag{25}$$

There is no decomposition of s.

Example 3. Finally, let us consider the series whose Hankel matrix is shown in Table 1.

The rank of this Hankel matrix is 6. We select the independent rows $\{L_1, L_{x_1}, L_{x_2}, L_{x_1x_2}, L_{x_2x_1x_2}, L_{x_1x_2x_1x_2}\}$ and the columns associated with the same words. This determinant has a maximal rank = 6.

The matrices $\mu(x_1)$ et $\mu(x_2)$ describe the action of the letters x_1 and x_2.

$$\mu(x_1) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{26}$$

and

$$\mu(x_2) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \tag{27}$$

The initial vector is

$$\gamma = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}^T \tag{28}$$

and the covector is

$$\lambda = \begin{pmatrix} 3 & 1 & 1 & 3 & 1 & 2 \end{pmatrix}. \tag{29}$$

By using the Jordan reduction on $\mu(x_1)$ (with Maple) we obtain

$$A = \mu(x_1)_P = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{30}$$

where the change of basis matrix is

$$P = \begin{pmatrix} 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 \\ -1 & -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 \end{pmatrix} \tag{31}$$

By this change of basis, $\mu(x_2)$ becomes

$$B = \mu(x_2)_P = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \tag{32}$$

In this new basis

$$\lambda_P = \begin{pmatrix} 0 & 1 & 1 & 0 & -1 & -1 \end{pmatrix} \tag{33}$$

and

$$\gamma_P = \begin{pmatrix} 0 & 1 & 1 & 0 & -1 & 0 \end{pmatrix}^T \tag{34}$$

In this case, we are lucky and the matrices A and B corresponding to $\mu(x_1)_P$ and $\mu(x_2)_P$ in the same basis directly present 3 diagonal blocks :

- the upper left block of size 2 corresponding to the series $s_1 = \dfrac{1}{1 - x_1x_2}$,
- the middle block of size 1 corresponding to the series $s_2 = \dfrac{1}{1 - (x_1 + x_2)}$,
- the lower right block of size 3 corresponding to the polynomial $s_3 = 1 + x_1x_2$. This last block is associated to a nilpotent representation.

5 AN APPLICATION TO DYNAMICAL SYSTEMS

Definition 5.1. *A bilinear dynamical system is a system of ordinary differential equations of the form*

$$\begin{cases} \dot{\mathbf{q}}(t) = \left(M_0 + \sum_{i=1}^{m} u_i(t) M_i \right) \mathbf{q}(t) \\ s(t) = \lambda \cdot \mathbf{q}(t), \end{cases} \tag{35}$$

where

Table 1: Hankel matrix of example 3.

	1	x_1	x_2	x_1^2	x_1x_2	x_2x_1	x_2^2	x_1^3	$x_1^2x_2$	$x_1x_2x_1$	$x_1x_2^2$	$x_2x_1^2$	$x_2x_1x_2$	$x_2^2x_1$	$x_2^3\cdots$
1	3	1	1	1	3	1	1	1	1	1	1	1	1	1	$1\cdots$
x_1	1	1	3	1	1	1	1	1	1	1	1	1	2	1	$1\cdots$
x_2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	$1\cdots$
x_1^2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	$1\cdots$
x_1x_2	3	1	1	1	2	1	1	1	1	1	1	1	1	1	$1\cdots$
x_2x_1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	$1\cdots$
x_2^2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	$1\cdots$
x_1^3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	$1\cdots$
$x_1^2x_2$	1	1	1	1	1	1	1	1	1	1	1	1	1	1	$1\cdots$
$x_1x_2x_1$	1	1	2	1	1	1	1	1	1	1	1	1	2	1	$1\cdots$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$

1. $\mathbf{u}(t) = (u_1(t),\ldots,u_n(t)) \in \mathbb{R}^n$ *is the (partwise continuous) input vector,*
2. $\mathbf{q}(t) \in \mathcal{M}$ *is the current state, where* $\mathcal{M}$ *is a real differential manifold, usually* $\mathbb{R}^m$*,*
3. $s(t) \in \mathbb{R}$ *is the output function.*

Definition 5.2. *The generating series G of a bilinear dynamical system (Fliess, 1981) is a formal power series with the alphabet* $X = \{z_o, z_1, \ldots z_m\}$*, where* z_i *for* $j > 0$ *correspond to the input* $u_i(t)$ *whereas* z_0 *corresponds to the drift. It is defined by*

$$\langle G|z_{j_0}\cdots z_{j_k}\rangle = \lambda\cdot M_{j_0}\cdots M_{j_k}\cdot \mathbf{q}(0). \qquad (36)$$

Theorem 5.1. *The generating series of bilinear dynamical system are rational. Inversely, every rational series is a generating series of a bilinear dynamical system.*

Proof. We take μ such that $\mu(z_i) = M_i$ for $i \geq 0$ and we denote $\gamma = \mathbf{q}(0)$. It follows directly that $\langle \lambda, \mu, \gamma\rangle$ is a rational series. □

Definition 5.3. *The Chen series measures the input contribution (Chen, 1971), and is independent of the system. The coefficients of the Chen series are calculated recursively by integration using the following two relations :*

- $\langle \mathcal{C}_u(t)|1\rangle = 1,$
- $\langle \mathcal{C}_u(t)|w\rangle = \int_0^t \langle \mathcal{C}_u(\tau)|v\rangle u_j(\tau)d\tau$ *for a word* $w = z_jv$*.*

The causal functional $y(t)$ is then obtained locally as the product of the generating series and the Chen series :

$$y(t) = \langle G||\mathcal{C}_u(t)\rangle = \sum_{w\in X^*} \langle G|w\rangle\langle \mathcal{C}_u(t)|w\rangle \qquad (37)$$

This formula is known as the *Peano–Baker formula*, as well as the *Fliess' fundamental formula*.

Now we apply the decomposition in the 3 above examples to the corresponding dynamical systems (identifying z_0 with x_1 and z_1 with x_2).

Example 1. The corresponding dynamical system is

$$\begin{cases} y_1'(t) = y_1(t) + u(t)y_2(t), & y_1(0) = 1,\\ y_2'(t) = y_2(t) + u(t)y_1(t), & y_2(0) = 0,\\ s(t) = y_2(t). \end{cases} \qquad (38)$$

Maple gives its solution is some complicated form. However using our decomposition into two dynamical systems

$$\overline{y}_1'(t) = \overline{y}_1(t)(1+u(t)),\quad \overline{y}_1(0) = \frac{1}{2},\quad s_1(t) = \overline{y}_1(t) \qquad (39)$$

and

$$\overline{y}_2'(t) = \overline{y}_2(t)(1-u(t)),\quad \overline{y}_2(0) = \frac{1}{2},\quad s_2(t) = -\overline{y}_2(t) \qquad (40)$$

we can easily obtain that

$$\begin{aligned} s(t) &= s_1(t) + s_2(t) = \\ &= \frac{1}{2}\left(exp\int_0^t (1+u(\tau))d\tau) - exp\int_0^t (1-u(\tau))d\tau\right) \end{aligned} \qquad (41)$$

Example 2. The corresponding dynamical system is

$$\begin{cases} y_1'(t) = u(t)(y_1(t) + y_2(t)), & y_1(0) = 1,\\ y_2'(t) = y_1(t) + y_2(t), & y_2(0) = 0,\\ s(t) = y_2(t). \end{cases} \qquad (42)$$

We can compute its solution directly

$$s(t) = \int_0^t exp\left(\int_0^{\tau_1} (1+u(\tau_2)d\tau_2\right)d\tau_1. \qquad (43)$$

$s(t)$ cannot be decomposed as a sum of two simpler expressions.

Example 3. The corresponding dynamical system cannot be solved directly. However, using the above decomposition we obtain $s(t) = s_1(t) + s_2(t) + s_3(t)$, where

$$\begin{aligned} s_1(t) =& 1 + \int_0^t \int_0^{\tau_1} u(\tau_2) d\tau_2 d\tau_1 + \\ & \int_0^t \int_0^{\tau_1} u(\tau_2) \int_0^{\tau_2} \int_0^{\tau_3} u(\tau_4) d\tau_4 d\tau_3 d\tau_2 d\tau_1 + \cdots \end{aligned} \tag{44}$$

corresponds to the first dynamical system and is the solution of the system

$$\begin{cases} y_1'(t) = s_2(t), & y_1(0) = 0, \\ y_2'(t) = u(t)s_1(t), & y_2(0) = 1, \\ s_1(t) = y_2(t). & \end{cases} \tag{45}$$

whereas

$$s_2(t) = exp\left(\int_0^t (1 + u(\tau)) d\tau\right) \tag{46}$$

corresponds to the second dynamical system and

$$s_3(t) = 1 + \int_0^t \int_0^{\tau_1} u(\tau_2) d\tau_2 d\tau_1 \tag{47}$$

is the solution of the third system.

6 CONCLUSIONS

In this paper, we presented an approach to the problem of decomposition of rational series in noncommutative variables into some simple series. The study of the simultaneous block–diagonalization has yet to be improved. We present an application of this decomposition to dynamical systems.

There are numerous further applications of this decomposition to dynamical systems and automata :

- The study of the stability of bilinear systems can be approached by using its generating series (Benmakrouha and Hespel, 2007) : in some cases, the output can be explicitly computed or bounded. The decomposition of this series into simple series would simplify this study in the other cases.
- In a bilinear system, the dependence or the independence of subsystems can be studied via the decomposition of the generating series of the system.
- A finite weighted automaton being another representation of a rational series, the property of decomposition of a rational series into simpler series is transferred to the corresponding finite weighted automaton. So we can define a simpler finite weighted automaton.

REFERENCES

Benmakrouha F. and Hespel C. (2007). Generating formal power series and stability of bilinear systems, *8th Hellenic European Conference on Computer Mathematics and its Applications (HERCMA 2007).*

Berstel J. and Reutenauer C. (1988). Rational series and their languages, Springer–Verlag.

Chen K.-T. (1971). Algebras of iterated path integrals and fundamental groups, *Trans. Am. Math. Soc.*, 156, 359–379.

Fevotte C. and Theis F.J. (2007). Orthonormal approximate joint block–diagonalization, technical report, Telecom Paris.

Fliess M. (1972). Sur certaines familles de séries formelles, Thèse d'état, Université de Paris–7.

Fliess M. (1974) Matrices de Hankel, *J. Maths. Pur. Appl.*, 53, 197-222.

Fliess M. (1976). Un outil algébrique : les séries formelles non commutatives, in "Mathematical System Theory" (G. Marchesini and S.K. Mitter Eds.), Lecture Notes Econom. Math. Syst., Springer Verlag, 131, 122-148.

Fliess M. (1981). Fonctionnelles causales non linéaires et indéterminées non commutatives, *Bull. Soc. Math. France*, 109, 3–40.

Gantmacher F.R. (1966). Théorie des matrices, Dunod.

Hespel C. (1998). Une étude des séries formelles non commutatives pour l'Approximation et l'Identification des systèmes dynamiques, Thèse d'état, Université de Lille–1.

Hespel C. and Martig C. (2006). Noncommutative computing and rational approximation of multivariate series, Transgressive Computing 2006, 271-286.

Jacob G. (1980) Réalisation des systèmes réguliers (ou bilinéaires) et séries génératrices non commutatives, Séminaire d'Aussois, RCP567, Outils et modèles mathématiques pour l'Automatique, l'Analyse des Systèmes, et le traitement du Signal.

Kaplanski I. (1969). Fields and rings, The University of Chicago Press, Chicago.

Reutenauer C. (1980). Séries formelles et algèbres syntactiques, *J. Algebra*, 66, 448-483.

Salomaa A. and Soittola M. (1978). Automata Theoretic aspects of Formal Power Series, Springer.

Schützenberger M.P (1961). On the definition of a family of automata, *Inform. and Control*, 4, 245-270.

ANALYSIS AND DESIGN OF COMPUTER ARCHITECTURE CIRCUITS WITH CONTROLLABLE DELAY LINE

N. V. Kuznetsov, G. A. Leonov, S. M. Seledzhi
Saint-Petersburg State University, Universitetski pr. 28, Saint-Petersburg, 198504, Russia
leonov@math.spbu.ru

P. Neittaanmäki
University of Jyväskylä, P.O. Box 35 (Agora), FIN-40014, Finland
pn@mit.jyu.fi

Keywords: Delay line, Phase locked loop, Nonlinear analysis, Clocked circuit.

Abstract: In this work classical and modern control theory methods are applied for rigorous mathematical analysis and design of different computer architecture circuits such as clock generators, synchronization systems and others. The present work is devoted to the questions of analysis and synthesis of feedback systems, in which there are controllable delay lines. In the work it is mathematically strictly shown that *RC*-chain can be used as a controllable delay line for different problems of circuit engineering if the chain is sequentially connected with hysteretic relay. This relay is either artificially introduced or shows itself as non-ideality of logic elements. The possibility of phase-locked loop application for time delay control is considered.

1 INTRODUCTION

The work is devoted to the questions of analysis and synthesis of feedback systems, in which there are controllable delay lines. First of all this is a class of controllable clock generators and clocked circuits, which perform the functions of summators (Cormen et al., 1990).

In clocked circuits it is necessary that the delay was by the one tact. For this purpose we need in a special setting of parameters of delay lines, which will be described in details. The generators, constructed on logic elements and delay lines, are not high-stable with respect to frequency (Ugrumov, 2000). Therefore, for their stabilization and synchronization by phase-locked loops it is necessary to introduce a controllable parameter in delay line. A class of such delay lines, the block-scheme of which is shown in Fig. 1, is considered.

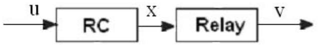


Figure 1: Delay line.

The *RC*-chains are often used in circuit engineering as delay lines (Ugrumov, 2000). We assume that the relation between the input u and the output x is described by the following standard equation of *RC*-chain

$$RC\frac{dx}{dt} + x = u(t), \qquad (1)$$

where R is a resistance, C is a circuit capacitance.

The relation between the input x and the output v is described by the graph of "relay with hysteresis"function, which is shown in Fig. 2. Here μ_1 and μ_2 are certain numbers from the interval $(0,1)$. The theory and practice of application of such relay blocks in feedback systems is well described in (Popov, 1979; Krasnosel'skii and Pokrovskii, 1983).

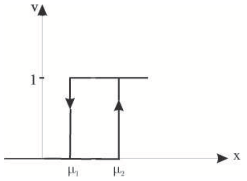


Figure 2: Relay with hysteresis.

In the present work we consider only the functions $u(t)$, which takes the values either 0 or 1 on certain intervals. Therefore, the solutions $x(t)$ of equation (1) are continuous, piecewise-differentiable and piecewise-monotone functions. It follows that the graph in Fig. 2 correctly defines the output $v(t)$. Further it will be shown that the hysteretic effect is of great importance for synthesis both of clock generators and of clocked summators. This effect always occurs in real (non-ideal) logic elements. Since the output of delay line is often the input of logic element, it is convenient to connect such hysteretic effect with *RC*-chain and to consider it in the frame of block-scheme in Fig. 1. In some cases for improvement of a quality of delay line operation it is possible to introduce additional block "relay with hysteresis", which provides a required delay time and stability of system operation.

We can show here the analogy with a classical study of Watt's regulator by I.A.Vyshnegradskii (Andronov and Voznesenskii, 1949; Leonov, 2001). Recall a main conclusion of Vyshnegradskii: "without friction the regulator is lacking". But if a friction "is not sufficient", then it is possible to introduce a special correcting device, dashpot, which provides a stable operation of system. In the case now being considered the friction is replaced by hysteretic effect and the above classical scheme of reasoning is repeated. This becomes especially clear if we consider the synthesis of clock generators.

For clocked summators it turns out rational the introduction of two-stage delay lines, which shift a unit impulse for the one tact. The latter permits us to use a three-bit summator for any summation, confining our attention to a minimal number of circuit elements.

The application of methods and technique of the classical control theory (Burkin et al., 1996; Leonov et al., 1996, Popov, 1979; Krasnosel'skii and Pokrovskii, 1983; Andronov and Voznesenskii, 1949) permits us to find the solution of considered problems, applying very simple mathematical constructions.

2 DELAY LINES FOR SYNTHESIS OF CONTROLLABLE CLOCK GENERATORS

Consider the block-scheme in Fig. 3 and, recall the table for Block AND-NOT output

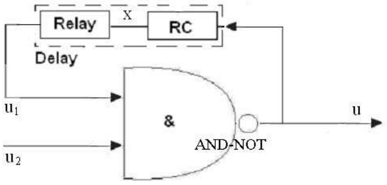


Figure 3: Clock generator on Block AND-NOT and delay line.

u_1	u_2	u
0	0	1
0	1	1
1	0	1
1	1	0

Truth table of Block AND – NOT

Let $u_2(t) = 0$ for $t < T$, $T > 0$. Then $u(t) = 1$ for $t < T$ and at the input $x(t)$ there occurs (after a transient process) the signal $x(t) = 1$. Suppose, $x(t) = 1$ on $[0,T]$. Then $u_1(t) = 1$ on $[0,T]$ and a system is in equilibrium:

$$1 = u_1(t) = x(t) = u(t), \quad u_2(t) = 0.$$

The inclusion of clock generator is realized by the change of u_2 from the state 0 to the state 1: $u_2(t) = 1$, $\forall t > T$. Then on the certain interval (T,T_1) we have $u(t) = 0$. This implies that $u_1(t) = 1$ for $t \in (T,T_1)$, where

$$T_1 = T + RC\ln\frac{1}{\mu_1} \qquad (2)$$

and $u_1(t) = 0$ on a certain interval (T_1,T_2).

Really, from equation (1) it follows that on (T,T_1) we have $x(t) = e^{-\alpha t}$, $\alpha = 1/RC$. In this case $u_1(t) = 1$ for $t \in (T,T_1)$, where T_1 is from relation (2), and $u_1(t) = 0$ for $t \in (T_1,T_2)$, where T_2 will be determined below. From the latter relation it should be that $u(t) = 1$ for $t \in (T_1,T_2)$. This implies the following relation

$$T_2 = T_1 + RC\ln\tfrac{1-\mu_1}{1-\mu_2}, \quad x(T_2) = \mu_2.$$

In the case when $\mu_1 = 1-\mu_2$, $\mu_2 \in (1/2,1)$, we obtain

$$\tau = T_1 - T_0 = T_2 - T_1 = RC\ln\tfrac{\mu_2}{1-\mu_2},$$
$$T_0 = T + RC\ln\tfrac{1}{\mu_2},$$

and 2τ-periodic sequence at the output u:

$$u(t) = 0, \quad \forall t \in [T_0, T_0+\tau),$$
$$u(t) = 1, \quad \forall t \in [T_0+\tau, T_0+2\tau).$$

Thus, the block-scheme in Fig. 3 is a clock generator with the frequency

$$\omega = \frac{1}{2\tau} = \left(2R\ln\frac{\mu_2}{1-\mu_2}\right)^{-1} C^{-1}. \qquad (3)$$

We compare this frequency with the frequency of harmonic *LC*-oscillator:

$$\omega = 1/\sqrt{LC} \qquad (4)$$

At present it is developed different methods of control of a frequency of harmonic oscillators by means of a slow (with respect to the high frequency ω) change of parameter C. It is especially widely extended the phase-locked loops (Viterbi, 1966; Lindsey, 1972). In the past decade similar constructions are actively developed and applied to the clock generators with frequency (3) (Solonina et al., 2000).

3 DELAY LINES FOR CLOCK IMPULSES

Consider the delay line, the block-scheme of which is shown in Fig. 1. Let $u(t)$ be 2τ-periodic sequence of impulses:

$$u(t) = 0, \forall t \in [0,\tau),\ u(t) = 1, \forall t \in [\tau, 2\tau). \qquad (5)$$

If we choose the initial data $x(0,x_0) = x_0$ so that the relation

$$\tau = RC \ln \frac{x_0}{1-x_0}, \quad x_0 \in (1/2,1), \qquad (6)$$

is satisfied, then $x(\tau,x_0) = 1-x_0, \quad x(2\tau,x_0) = x_0$. In this case the graph of 2τ-periodic function $x(t)$ is shown in Fig. 4.

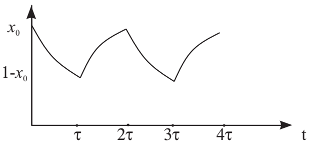


Figure 4: Periodic output of RC-chain.

It is well known (Leonov, 2001) that for all other solutions of equation (1) $x(t,y_0)$ the following relation

$$\lim_{t\to+\infty} (x(t,x_0) - x(t,y_0)) = 0 \qquad (7)$$

is satisfied. If we choose $x_0 > \mu_2$, $1-x_0 < \mu_1$, then relation (7) implies that after transient process, at the output v (of delay line) we obtain 2τ-periodic sequence of impulses:

$$\begin{aligned} v(t) &= 0,\ \forall t \in \left[RC\ln\tfrac{x_0}{\mu_1}, \tau + RC\ln\tfrac{x_0}{1-\mu_2}\right), \\ v(t) &= 1,\ \forall t \in \left[\tau + RC\ln\tfrac{x_0}{1-\mu_2}, 2\tau + RC\ln\tfrac{x_0}{\mu_1}\right). \end{aligned} \qquad (8)$$

Note that for $\mu_1 = 1-x_0+\varepsilon$, $\mu_2 = x_0-\varepsilon$, where $\varepsilon > 0$ is a small parameter, from (8) we have

$$\begin{aligned} v(t) &= 0, \quad \forall t \in [\tau_\varepsilon, \tau+\tau_\varepsilon), \\ v(t) &= 1, \quad \forall t \in [\tau+\tau_\varepsilon, 2\tau+\tau_\varepsilon), \end{aligned} \qquad (9)$$

where

$$\tau_\varepsilon = RC\ln\left(\frac{x_0}{1-x_0+\varepsilon}\right) \xrightarrow[\varepsilon\to 0]{} \tau. \qquad (10)$$

Recall that $x_0 \in (1/2,1)$ and τ is determined from relation (6).

Thus, the block-scheme in Fig. 1 realizes asymptotically the time delay τ: after transient process (see relation (7)) at the output v we observe relation (9), in which case relation (10) is satisfied.

Consider now a certain extension of the above case. Let $u(t)$ be a certain sequence of clock impulses (not necessarily 2τ-periodic) such that

$$u(t) = 0, \quad \forall t \in [2k\tau, (2k+1)\tau),\ k = 0,1,\ldots$$

and on each of intervals $((2k+1)\tau, 2k+2)\tau)$ it can take the value either 0 or 1.

Now we consider the case when the delay line operates in working conditions after transient process. In this case, taking into account the above reasoning, we can assume that for the certain fixed k there occur the following restrictions:

$$\begin{aligned} &u(t) = 1, \quad \forall t \in [(2k+1)\tau, 2(k+1)\tau) \\ &x((2k+1)\tau) \in (0, 1-x_0), \end{aligned}$$

where x_0 satisfies relation (6).

We shall show that in this case it can be made such a choice of parameters of delay line, for which asymptotically (at $\varepsilon \to 0$) the delay time of unit impulse is τ. For this purpose we can take the obvious inequalities

$$\begin{aligned} &x(t,(2k+1)\tau,0) \le x(t,(2k+1)\tau, x((2k+1)\tau) \le \\ &\le x(t,(2k+1)\tau, 1-x_0), \quad \forall t \ge (2k+1)\tau. \end{aligned}$$

Here $x((2k+1)\tau,(2k+1)\tau,y_0) = y_0$. By the previous relations $\mu_1 = 1-x_0+\varepsilon$, $\mu_2 = x_0-\varepsilon$ we obtain

$$\begin{aligned} &v(t) = 0,\ \forall t \in ((2k+1)\tau, (2k+1)\tau + \tau_\varepsilon), \\ &v(t) = 1,\ \forall t \in ((2k+1)\tau + \widetilde{\tau}_\varepsilon, (2k+1)\tau + \tau_\varepsilon + \widetilde{\widetilde{\tau}}_\varepsilon). \end{aligned}$$

Here

$$\widetilde{\tau}_\varepsilon = RC\ln(\frac{1}{1-x_0+\varepsilon}), \widetilde{\widetilde{\tau}}_\varepsilon = RC\ln(\frac{x_0-\varepsilon}{1-x_0+\varepsilon}).$$

Choosing $x_0 = 1-\sqrt{\varepsilon}$, we obtain the following formulas for parameters of delay line, which shifts unit impulse with accuracy up to $\sqrt{\varepsilon}$ for time τ:

$$\mu_1 = \sqrt{\varepsilon}+\varepsilon, \mu_2 = 1-\mu_1, RC = \tau/\ln\frac{1}{\sqrt{\varepsilon}}. \qquad (11)$$

This implies that for the asymptotical shift of unit impulse for time 2τ it is necessary to apply two-stage delay line with parameters (11) (Fig. 5). We proceed now to the clocked circuits for bit summation (Cormen et al., 1990) (Fig. 6). Here Σ is a standard summator, at the input of which we have three bits, $c_0 = 0$. As the delay line we can use a two-stage delay line with parameters (11) (Fig. 5). The time between the

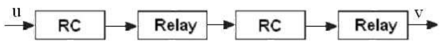


Figure 5: Two-stage delay line.

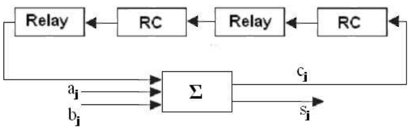


Figure 6: Clocked summator.

arrival of the signals a_j and a_{j+1} (and also b_j and b_{j+1}) is equal to τ. It is easily seen that the output $s_k s_{k-1} \dots s_0$ is a sum of two numbers $a_{k-1} a_{k-2} \dots a_0$ and $b_{k-1} b_{k-2} \dots b_0$. Thus, the delay line considered permits us to construct the summators with minimal number of circuit elements.

4 CONCLUSIONS

In the present work it is mathematically rigorously shown that *RC*-chain can be used as a controllable delay line for different problems of circuit engineering if the chain is sequentially connected with hysteretic relay. This relay is either artificially introduced or shows itself as non-ideality of logic elements.

ACKNOWLEDGEMENTS

The work was partly supported by the projects RFFI 07-01-00151, LSS-2387.2008.1, AVZP-2.1.1/3889.

REFERENCES

Andronov, A.A., Voznesenskii, I.N., 1949. *About the works of D.K. Maxwell, I.A. Vyshnegradskii and A. Stodola in the field of control theory*. M.: Izd. AN USSR.

Burkin, I.M., Leonov, G.A., Shepeljavy, A.I., 1996. *Frequency Methods in Oscillation Theory*. Dordrecht: Kluwer.

Cormen, T.H., Leiserson, C.E., Rivest, R.L., 1990. *Introduction to Algorithms*. Cambridge, Massachusetts: MIT Press.

Horowitz, P., Hill, W., 1998. *The Art of Electronics*. Cambridge Univ. Press.

Krasnosel'skii, M.A. and Pokrovskii, A.V., 1983. *Systems with hysteresis*. M.: Nauka.

Kuznetsov, N.V., 2008. *Stability and Oscillations of Dynamical Systems: Theory and Applications* Jyväskylä Univ. Press.

Leonov, G.A., 2006. *Phase synchronization. Theory and Applications* Automation and remote control, N 10, pp. 47–85. (survey)

Leonov, G.A., Seledzhi, S.M., 2005. *Design of Phase-Locked Loops for Digital Signal Processors*, Int. J. Innovative Computing, Information & Control. Vol.1, N4, pp. 1–11.

Leonov, G.A., 2001. *Mathematical problems of control theory*. World Scientific.

Leonov, G., Ponomarenko, D., Smirnova, V., 1996. *Frequency-Domain Methods for Nonlinear Analysis. Theory and Applications*. World Scientific.

Lindsey, W., 1972, *Synchronization systems in communication and control*, Prentice-Hall.

Popov, E.P., 1979. *The theory of nonlinear systems of automatic regulation and control*. M.: Nauka.

Solonina, A., Ulahovich, D., Jakovlev, L., 2000. *The Motorola Digital Signal Processors*. BHV, St. Petersburg. (in Russian)

Ugrumov, E., 2000. *Digital engineering*, BHV, St.Petersburg. (in Russian)

Viterbi, A., 1966. *Principles of coherent communications*, McGraw-Hill. New York.

SIMPLE DESIGN OF THE STATE OBSERVER FOR LINEAR TIME-VARYING SYSTEMS

Yasuhiko Mutoh
Department of Applied Science and Engineering, Sophia University, 7-1, Kioicho, Chiyoda-ku, Tokyo, Japan
y_mutou@sophia.ac.jp

Keywords: Pole placement, State observer, Linear time-varying system.

Abstract: A simple design method of the Luenberger observer for linear time-varying systems is proposed in this paper. The paper first propose the simple calculation method to derive the pole placement feedback gain vector for linear time-varying systems. For this purpose, it is shown that the pole placement controller can be derived simply by finding some particular ”output signal” such that the relative degree from the input to this output is equal to the order of the system. Using this fact, the feedback gain vector can be calculated directly from plant parameters without transforming the system into any standard form. Then, this method is applied to the design of the observer, i.e., because of the duality of linear time-varying system, the state observer can be derived by un-stabilization of the state error equation.

1 INTRODUCTION

The design of the pole placement and the state observer for linear time-varying systems is well established problem. As for the linear time-invariant case, if the system is controllable, the pole placement controller can be designed, and, if observable, the state observer can be designed. However, many of those design method need a complicated calculation procedure. In this paper, a simple design method of the Luenberger observer for linear time varying systems is proposed.

Since, the observer design problem is the dual problem of the pole placement, simplified calculation method to derive the pole placement feedback gain vector for linear time-varying systems should be considered first. Usually, the pole placement procedure needs the change of variable to the Flobenius standard form, and hence, is very complicated (e.g., Michael Valášek and Nejat Olgaç). To simplify this procedure, it will be shown that the pole placement controller can be derived simply by finding some particular ”output signals” such that the relative degree from the input to this output is equal to the order of the system. This is motivated from the fact that the input-output linearization of a certain type of nonlinear systems is equivalent to the entire state linearization, if the relative degree of the system is equal to the system order. Using this fact, the feedback gain vector can be calculated directly from plant parameters without transforming the system into any standard form.

Because of the duality of the linear time-varying system, the state observer can be derived by unstabilizing the state error eqation. This implies that the simplified pole placement technique can be applied to the design of the state observer for linear time-varying systems to obtain simpler design method than existing methods from the point of view of the calculational compexity.

In the sequel, the simple pole placement technique is proposed in Section 2, and then, this method is used to the observer design problem in Section 3.

2 POLE PLACEMENT OF LINEAR TIME-VARYING SYSTEMS

Consider the following linear time-varying system with a single input.

$$\dot{x} = A(t)x + b(t)u \tag{1}$$

Here, $x \in R^n$ and $u \in R^1$ are the state variable and the input signal respectively. $A(t) \in R^{n\times n}$ and $b(t) \in R^n$ are time-varying parameter matrices. The problem is to find the state feedback

$$u = k^T(t)x \tag{2}$$

which makes the closed loop system equivalent to the time invariant linear system with arbitrarily stable poles.

Now, consider the problem of finding a new output signal $y(t)$ such that the relative degree from u to y is n. Here, $y(t)$ has the following form.

$$y(t) = c^T(t)x(t) \tag{3}$$

Then, the problem is to find a vector $c(t) \in R^n$ that satisfies this condition.

Lemma 1. Let $c_k^T(t)$ be defined by the following equation.

$$c_k^T(t) = \dot{c}_{k-1}^T(t) + c_{k-1}^T(t)A(t), \quad c_0^T(t) = c^T(t) \tag{4}$$

The relative degree from u to y defined by (3) is n, if and only if

$$\begin{aligned} c_0^T(t)b(t) &= c_1^T(t)b(t) = \cdots = c_{n-2}^T(t)b(t) = 0 \\ c_{n-1}^T(t)b(t) &= 1 \end{aligned} \tag{5}$$

(Here, $c_{n-1}^T(t)b(t) = 1$ without loss of generality.)
Proof : By differentiating y successively using (5), the following equations are obtained from (1) and (3).

$$\begin{aligned} y &= c^T(t)x \\ &= c_0^T(t)x \\ \dot{y} &= \left(\dot{c}^T(t) + c^T(t)A(t)\right)x + c^T(t)b(t)u \\ &= c_1^T(t)x + c_0^T(t)b(t)u \\ &= c_1^T(t)x \\ \ddot{y} &= \left(\dot{c}_1^T(t) + c_1^T(t)A(t)\right)x + c_1^T(t)b(t)u \\ &= c_2^T(t)x + c_1^T(t)b(t)u \\ &= c_2^T(t)x \qquad (6) \\ &\vdots \\ y^{(n-1)} &= c_{n-1}^T(t)x + c_{n-2}^T(t)b(t)u \\ &= c_{n-1}^T(t)x \\ y^{(n)} &= c_n^T(t)x + c_{n-1}^T(t)b(t)u \\ &= c_n^T(t)x + u \end{aligned}$$

This implies that the relative degree from u to y is n. ∇∇

Lemma 2. If $c^T(t)$ satisfies the condition that the relative degree from u to y is n, then we have the following equation.

$$\begin{aligned} &[c_0^T(t)b(t),\ c_1^T(t)b(t),\ \cdots,\ c_{n-1}^T(t)b(t)] \\ &\quad = [c^T(t)b_0(t),\ c^T(t)b_1(t),\ \cdots,\ c^T(t)b_{n-1}(t)] \end{aligned} \tag{7}$$

where $b_i(t)$ is defined by

$$b_i(t) = A(t)b_{i-1}(t) - \dot{b}_{i-1}(t), \quad b_0(t) = b(t) \tag{8}$$

Proof : First, the following is trivial.

$$c_0^T(t)b(t) = c^T(t)b(t) = c^T(t)b_0(t) \tag{9}$$

From (5), we have

$$\dot{c}_0^T(t)b(t) = -c_0^T(t)\dot{b}(t) \tag{10}$$

which implies

$$\begin{aligned} c_1^T(t)b(t) &= \dot{c}_0^T(t)b(t) + c_0^T(t)A(t)b(t) \\ &= -c_0^T(t)\dot{b}(t) + c_0^T(t)A(t)b(t) \\ &= c_0^T(t)b_1(t) \\ &= c^T(t)b_1(t) \end{aligned} \tag{11}$$

In a similar fashion, from (5) and (11), we have

$$\begin{aligned} \dot{c}_0^T(t)b_1(t) &= -c_0^T(t)\dot{b}_1(t) \\ \dot{c}_1^T(t)b(t) &= -c_1^T(t)\dot{b}(t) \end{aligned} \tag{12}$$

which implies

$$\begin{aligned} c_2^T(t)b(t) &= \dot{c}_1^T(t)b(t) + c_1^T(t)A(t)b(t) \\ &= -c_1^T(t)\dot{b}(t) + c_1^T(t)A(t)b(t) \\ &= c_1^T(t)b_1(t) \\ &= \dot{c}_0^T(t)b_1(t) + c_0^T(t)A(t)b_1(t) \\ &= -c_0^T(t)\dot{b}_1(t) + c_0^T(t)A(t)b_1(t) \\ &= c_0^T(t)b_2(t) \\ &= c^T(t)b_2(t) \end{aligned} \tag{13}$$

By continuing the same process, (7) is derived. ∇∇

From Lemma 2, (5) implies

$$\begin{aligned} &[c_0^T(t)b(t), c_1^T(t)b(t), \cdots,\ c_{n-1}^T(t)b(t)] \\ &\quad = [c^T(t)b_0(t),\ c^T(t)b_1(t),\ \cdots,\ c^T(t)b_{n-1}(t)] \\ &\quad = c^T(t)[b_0(t),\ b_1(t),\ \cdots,\ b_{n-1}(t)] \\ &\quad = c^T(t)U_c(t) \\ &\quad = [0,\ 0,\ \cdots,\ 1] \end{aligned} \tag{14}$$

Here,

$$U_c(t) = [b_0(t),\ b_1(t),\ \cdots,\ b_{n-1}(t)] \tag{15}$$

where, $U_c(t)$ the controllability matrix for linear time-varying system (1). If $U_c(t)$ is nonsingular for all $t \in [0, \infty)$, the system is said to be controllable. Hence, we have the following Theorem.

Theorem 1. If the system (1) is controllable, there exists a vector $c(t)$ such that the relative degree from u to $y = c^T(t)x$ is n. And, such a vector, $c(t)$ is given by

$$c^T(t) = [0,\ 0,\ \cdots,\ 1]U_c^{-1}(t) \tag{16}$$

∇∇

The next step is to derive the state feedback for the arbitrary pole placement. Let $q(p)$ be a desired stable polynomial of the differential operator, p, i.e.,

$$q(p) = p^n + \alpha_{n-1}p^{n-1} + \cdots + \alpha_0 \tag{17}$$

By multiplying $y^{(i)}$ by α_i $(i = 0, \cdots, n-1)$ and then summing them up, the following equation is obtained, using (5) and (6).

$$q(p)y = d^T(t)x + u \tag{18}$$

where $d(t) \in R^n$ is defined by the following.

$$d^T(t) = [\alpha_0, \alpha_1, \cdots, \alpha_{n-1}, 1]\begin{bmatrix} c_0^T(t) \\ c_1^T(t) \\ \vdots \\ c_{n-1}^T(t) \\ c_n^T(t) \end{bmatrix} \tag{19}$$

Hence, the state feedback,

$$u = -d^T(t)x + r \tag{20}$$

makes the closed loop system as follows.

$$q(p)y = r \tag{21}$$

where r is an external input signal. This method is regarded as an extension of Ackermann's pole placement method to the time-varying case.

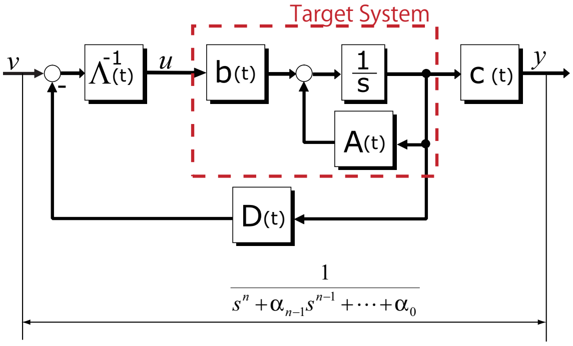


Figure 1: Blockdiagram of Pole Placement for a Linear Time-Varying System.

This control system can be summarized as follows. The given system is

$$\dot{x} = A(t)x + b(t)u \tag{22}$$

and, using (16) and (19), the state feedback for the pole placement is given by

$$u = -d^T(t)x. \tag{23}$$

Then, the closed loop system becomes

$$\dot{x} = (A(t) - b(t)d^T(t))x. \tag{24}$$

At the same time, we have (21) as another representation of the closed loop system. This can be explained as follows.

Let $T(t)$ be the time varying matrix defined by

$$T(t) = \begin{bmatrix} c_0(t)^T \\ c_1(t)^T \\ \vdots \\ c_{n-1}^T(t) \end{bmatrix} \tag{25}$$

and define the new state variable w by

$$x = T(t)w, \qquad w = \begin{bmatrix} y(t) \\ \dot{y}(t) \\ \vdots \\ y^{(n-1)}(t) \end{bmatrix} \tag{26}$$

Theorem 2. If the system (1) is controllable, then, the matrix for the change of variable, $T(t)$, given by (25) is nonsingular for all t. $\nabla\nabla$

This theorem can be proved by simple calculation as for the time invariant case.

Then, (24) is transformed into

$$\begin{aligned} \dot{w} &= \{T(t)(A(t) - b(t)d^T(t))T^{-1}(t) - T(t)\dot{T}^{-1}(t)\}w \\ &= \begin{bmatrix} 0 & 1 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ \vdots & & & 1 \\ -\alpha_0 & \cdots & \cdots & -\alpha_{n-1} \end{bmatrix} w = A^* w \end{aligned} \tag{27}$$

This implies that the closed loop system is equivalent to the time invariant linear system which has the desired closed loop poles. $(\det(pI - A^*) = q(p))$

3 STATE OBSERVER

In this section, we consider the design of the observer for the following linear time-varying system.

$$\begin{aligned} \dot{x} &= A(t)x + b(t)u \\ y &= c^T(t)x \end{aligned} \tag{28}$$

Here, $y \in R$ is the output signal of this system. The problem is to design the full order state observer of (28). Consider the following system as a candidate of the observer.

$$\begin{aligned} \dot{z} &= F(t)z + b(t)u + h(t)y \\ &= F(t)z + b(t)u + h(t)c^T(t)x \end{aligned} \tag{29}$$

where $F(t) \in R^{n \times n}$, and $h(t) \in R^n$. Define the state error $e \in R^n$ by

$$e = x - z \tag{30}$$

Then, e satisfies the following error equation.

$$\dot{e} = F(t)e + (A(t) - F(t) - h(t)c^T(t))x \tag{31}$$

Hence, (29) is a state observer of (28) if $F(t)$ and $h(t)$ satisfy the following condition.

$$\begin{aligned} F(t) &= A(t) - h(t)c^T(t) \qquad (32)\\ F(t) &: \text{arbitrarily stable matrix} \end{aligned}$$

Consider the pole placement control problem of the following system.

$$\dot{x} = -A^T(t)x + c(t)u \qquad (33)$$

From the property of the duality of the time varying system, if the pair $(A(t), c^T(t))$ is observable, the pair $(-A^T(t), c(t))$ is controllable. This implies that if the system (28) is observable, there is a state feedback for the arbitrary pole placement for the system (33). Let $(\lambda_1, \lambda_2, \cdots, \lambda_n)$ be the set of desired stable closed loop poles.

Suppose that

$$u = k^T(t)x \qquad (34)$$

is the state feedback for (33) with the desired **unstable** closed loop poles, $(-\lambda_1, -\lambda_2, \cdots, -\lambda_n)$. The closed loop system is

$$\dot{x} = (-A^T(t) + c(t)k^T(t))x \qquad (35)$$

This implies that, using the appropriate change of variable, $x = P(t)w$, (35) can be transformed into the following time invariant system.

$$\begin{aligned} \dot{w} &= \{P^{-1}(t)(-A^T(t) + c(t)k^T(t))P(t) \\ &\qquad - P^{-1}(t)\dot{P}(t)\}w \\ &= -F^{*T}w \qquad (36) \end{aligned}$$

Here, the eigenvalues of $-F^*$ are $(-\lambda_1, -\lambda_2, \cdots, -\lambda_n)$.

It is also well known that if the fundamental matrices of (35) and its dual system,

$$\dot{x} = (A(t) - k(t)c^T(t))x \qquad (37)$$

are $\Phi(t, t_0)$ and $\Psi(t, t_0)$, respectively, then,

$$\Phi(t, t_0) = \Psi^T(t_0, t). \qquad (38)$$

Furthermore, by the change of variable,

$$x = (P^T(t))^{-1}\xi \qquad (39)$$

(35) is transformed into

$$\begin{aligned} \dot{\xi} &= (P^T(A(t) - k(t)c^T(t))(P^T)^{-1} - P^T(\dot{P}^T)^{-1})\xi \\ &= (P^T(A(t) - k(t)c^T(t))(P^T)^{-1} + \dot{P}^T(P^T)^{-1})\xi \\ &= \left(P^{-1}(A^T(t) - c(t)k^T(t))P + P^{-1}\dot{P}\right)^T \xi \\ &= F^*\xi \qquad (40) \end{aligned}$$

Hence, by choosing

$$h(t) = k(t) \qquad (41)$$

(29) becomes the observer for (28), and the state error equation becomes

$$\dot{e} = F(t)e \qquad (42)$$

which is equivalent to (40). That is, if the system (28) is observable, it is possible to design $h(t)$ so that the state estimation error equation is equivalent to the time invariant homogeneous system which has the arbitrary stable poles.

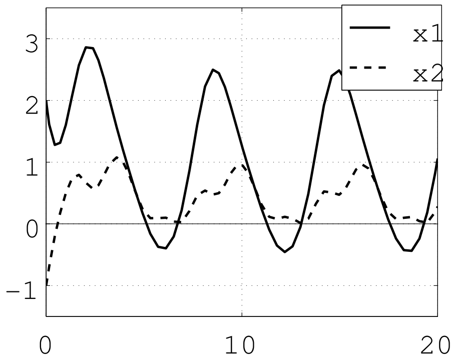


Figure 2: Responce of the state variable (x) of the system.

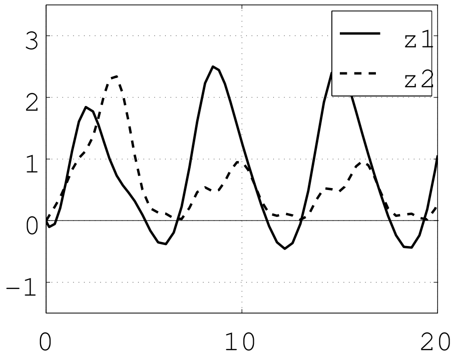


Figure 3: Responce of the state variable of the observer (z).

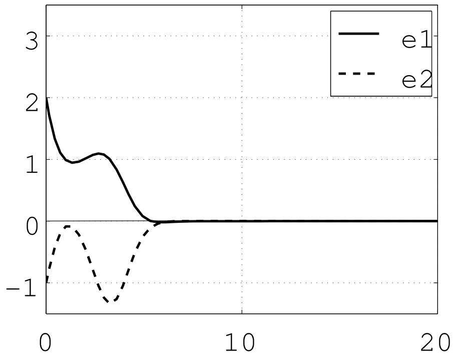


Figure 4: Responce of the state error ($e = x - z$).

Example 1. Consider the following system.

$$\begin{aligned} \dot{x} &= A(t)x + b(t)u \\ y &= c^T(t)x \qquad (43) \end{aligned}$$

where

$$\begin{aligned} A(t) &= \begin{bmatrix} -1, & 1 \\ -1+\sin 2t - \cos t, & -3+\cos t \end{bmatrix} \\ b(t) &= \begin{bmatrix} 2+\sin t \\ 0 \end{bmatrix} \\ c^T(t) &= \begin{bmatrix} 2+\sin 0.5t & 0 \end{bmatrix} \qquad (44) \end{aligned}$$

This is a stable time-varying observable system. Fig.2 shows the response of the state variable of this system with

$$u = \sin t \qquad (45)$$

The state observer is the following.

$$\dot{z} = (A(t) - h(t)c^T(t))z + b(t)u + h(t)y \quad (46)$$

where we choose the desired observer poles as -1 and -2. (The numerical details are omitted in this draft paper.)

Fig.3 and 4 show the state variable of the observer and the state error.

4 CONCLUSIONS

In this paper, one design method for the state observer for linear time-varying systems is proposed. We first proposed the simple calculation method for the pole placement state feedback gain for liner time-varying system. Feedback gain can be derived directly from the plant parameter without the transformation into any standard form.

In this method, since the transformation of the given system into the Flobenius standard form is not required, the design procedure is very simple. It was shown that if the system is observable, then the state observer can be obtained with arbitrarily stable observer poles.

REFERENCES

Charles C. NguyenC *Arbitrary eigenvalue assignments for linear time-varying multivariable control systems*. International Journal of Control, 45-3, 1051–1057, 1987

Chi-Tsong ChenC *Linear System Theory and Design (Third edition)*. OXFORD UNIVERSITY PRESS, 1999

T. KailathC *Linear Systems*. Prentice-Hall, 1980

Michael Valášek, Nejat Olgaç *Efficient Eigenvalue Assignment for General Linear MIMO systems*. Automatica, 31-11, 1605–1617, 1995

Michael Valášek, Nejat OlgaçC *Pole placement for linear time-varying non-lexicographically fixed MIMO systems*. Automatica, 35-1, 101–108, 1999

MODELING APPROACH FOR NETWORKED EMBEDDED SYSTEMS WITH HETEROGENEOUS COMMUNICATION
Modeling Common Gateway Functionalities for Interconnection

Johannes Klöckner, Marcus Müller, Wolfgang Fengler and Yixin Huang
Computer Architecture Group, Ilmenau University of Technology, P.O. Box 100565, Ilmenau, Germany
{johannes.kloeckner, marcus.mueller, wolfgang.fengler}@tu-ilmenau.de, huangyixin81@hotmail.com

Keywords: Model based design, Fieldbus, FlexRay, CAN, Building blocks, Network simulation, MLDesigner, Gateway.

Abstract: This paper presents a method to create system level models of gateway facilities to combine embedded fieldbus networks. Here automotive applications are used as representative with CAN and FlexRay. A strategy is introduced to create gateway models, which provide high flexibility in terms of reusability, replaceability, extensibility and flexibility. Todays systems often incorporate a very complex heterogeneous communication structure with real time requirements. These characteristics have an influence on the system design process. The goal of avoiding design errors in early development stages requires the validation on hierarchical functional models of various abstraction levels using simulation based analysis prior to hardware design. A current approach includes problem oriented system analysis. This approach is extended to analyze the system behavior of heterogeneous embedded systems focusing on switching mechanisms. Here lies the challenge on the connection of subsystems with distinct fieldbuses realizing different paradigms. Regarding real time aspects paradigm switches can be a source of errors in the system design.

1 INTRODUCTION

As a result of the increase in both system complexity and the amount of incorporated functionality a single embedded system has to be considered as a composition of several heterogeneous subsystems. Due to this the intra-system networking of an embedded system provides new challenges in system design. Developing such complex systems requires a lot of planning and design decisions on a profound knowledge of system behavior. The heterogeneity in a compound system stems from differences in hardware, software and subsystem architectures. In networked embedded systems the communication is an important aspect. Based on the complexity of systems and distributed applications the amount of communication between individual subsystems grows. Using communication technologies with different paradigms, e.g. event or time triggered, influence the system behavior.

The heterogeneity of large systems represents a great challenge in the system design. This high grade of variability complicates the development process. Mistakes in early system design stages can negatively influence the performance and development costs. Avoiding these errors requires the validation on functional models of various abstraction levels using simulation based analysis prior to hardware design.

To analyze the performance on system level (Henia et al., 2005) use an approach based on a scheduling analysis. A model based design approach (Salzwedel, 2004) enables an efficient top down development of a complete system using hierarchically composed building blocks, thus providing a high grade of reusability and exchangeability.

Aim of the presented work is the extension of an existing modeling approach studying networked embedded systems to allow the modeling of heterogeneous networked systems in a common way and the analysis of system behavior. It is based on the tool MLDesigner (MLDesign Technologies Inc., 2007) and fulfills the requirements of a model based design approach allowing system analysis and development. An example for a complex heterogeneous networked embedded system is an automobile. It is a system with a large amount of functionality, real time requirements and distributed characteristics. The complex communication infrastructure contains several protocols with different paradigms, e.g. FlexRay (FlexRay Consortium, 2005) and Controller Area Network (CAN). The mentioned properties can be found in a lot of systems belonging to areas engaged with automation and control. This paper is organized as follows. Section 2 gives a short description of the used modeling tool. Section 3 introduces related work and the basic modeling strategy. Section 4 describes the modeling concept. Section 5 presents the drawn conclusions and a brief overview of further development steps.

2 MODELING TOOL

Currently many tools exist, that realize model-based design to support the development process. In this work the tool MLDesigner by MLDesign Technologies, Inc. is used. The tool is dedicated to improving the design process from early concepts to implementation with mission and system level design. MLDesigner extends the Ptolemy project of UC Berkeley (The Ptolemy Project, 2009) with modeling paradigms. Different models of computation so called domains are provided, e.g. *discrete event domain* (DE) and *finite state machines* (FSM), *synchronous data flow domain* (SDF). In this work the *discrete event domain* and *finite state machines* are used. These two are well suited to modeling networked embedded systems.

Now a short introduction to the terms of MLDesigner. The *System* is the top level element in the modeling hierarchy. The building blocks can be atomic blocks called *Primitives*, specified as FSM or in C/C++ code, or hierarchical *Modules*. To communicate with their environment building blocks can use linked variables or ports. *Ports* are represented as arrows on the bounding box of a building block and are interconnected by signal paths. So called *Wormholes* allow the embedding of building blocks belonging to different domains.

3 RELATED WORK

In (Klöckner et al., 2008) a modeling strategy for networked embedded systems is discussed exemplified by a top-down developed generalized FlexRay protocol model. The communication system is defined as a composition of three different elements: *Host*, *Communication Controller* (CC) and *Channel* as shown in Figure 1. The combination of a host and a CC is called node. The corresponding protocol can be implemented within the CC, e.g. synchronization, error detection, message transmission and reception. In addition, the CC provides several services to the host to be interfaced by its application. The application of a node can be described within the host. The host uses CC provided services to configure the CC, initiate the sending and receiving operations and process the received data. According to this the host contains a specialized sublayer realizing the protocol related access to the CC. At this point the division into different functional layers is visible, the host describes the functionality and the CC describes the type of communication. Nodes are grouped to communication clusters by connecting them to a channel.

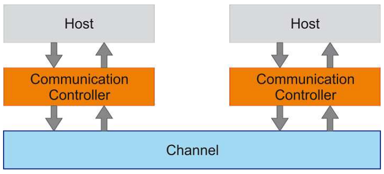


Figure 1: Basic Model Structure for networked embedded systems (Klöckner et al., 2008).

The physical characteristics of the connection between nodes is described and modeled within a channel. On this level the physical delay of a transmission is determined by the respective message data length and a fault injection model simulates the transmission errors caused by physical medium and environmental influences.

The focus of the approach (Klöckner et al., 2008) lies on homogeneous communication systems. The model structure intends the extension of the approach to allow the modeling of heterogeneous systems.

4 GATEWAY MODEL

In order to provide concepts to monitor both the behavior of networked systems and the communication across different types of networks with different specifications and protocols, a concept for a gateway model which provides high flexibility in terms of reusability, replaceability, extensibility and flexibility to connect different communication technologies in a common way, is needed. Based on the model structures presented in (Klöckner et al., 2008) (Müller, 2007) and the libraries developed in these works, the creation of a customizable gateway module is possible. This allows the combination of heterogeneous subsystems in order to analyze the overall system behavior. Regarding the system architecture presented in Figure 1 a gateway is a special type of node. Normally a node is a combination of a host and a CC. A node realizing gateway functionality is now a combination of several *hosts*, associated CCs and a *Gateway Core*. The basic gateway structure is shown in Figure 2. The CCs contain the identical functionality as described before. A common strategy to model an

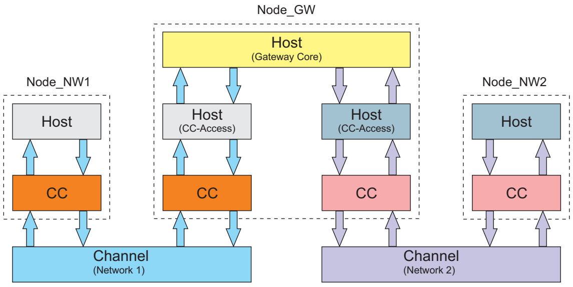


Figure 2: Extended Model Structure for networked embedded systems.

application is not yet designed. Therefore the hosts realize the specific sublayers to access the CCs. The *Gateway Core* can be seen as an application of the node providing the gateway base function. This concept fits the described modeling strategy. Anticipating a more detailed model containing different applications in a node divided in several tasks it is possible to extend this architecture to enable the analysis of the system behavior including other effects, e.g. properties of task scheduling.

Gateway facilities handle the data exchange between heterogeneous computer networks. Normally, they are designed to combine specified networks, and fulfill specified needs, which are based on the application. Hence, there is no predefined specification of a gateway. Although there are no fixed design approaches for gateway design, certain common functions can still be assumed for any specialized communications gateway: routing, protocol conversion, dataflow control, real time constraint checking for message transmission and message filtering or segmentation. The purpose of this paper is to develop a universal gateway model, which facilitates the communications between different networks. The design of the presented approach is similar to the idea of a time triggered gateway described in (Shaheen et al., 2007).

As described in (Fahmy, 1995), the gateway model is based on a shared medium approach, which uses an internal network to exchange uncommitted messages between *Host* and *Gateway Core*. Received frames are converted to an uncommitted protocol, after which the information is handled by the gateway module. Finally the information is converted to match the destination protocol. The uncommitted protocol contains three values: *Source Identifier*, *Destination Identifier* and *Datafield*. The routing information of each message can be identified by *Source Identifier* and *Destination Identifier*. The data of the received message is stored in the *Datafield* of the uncommitted message. To fulfill the flexibility requirement the gateway model is a composition of two kinds of modules: protocol related and unrelated modules. Only uncommitted gateway messages can be processed inside the protocol unrelated modules, and the protocol related modules process the specific protocol messages. In order to extend the gateway model to support a larger amount of protocols, only the protocol related modules need to be created, while the protocol unrelated modules remain unchanged. This approach provides enormous reusability, replace ability and extensibility. It allows an easy generation of a customized gateway to realize relevant system characteristics.

An important aspect of the gateway functionality refers to (Hörner, 2007). A gateway should provide at least two methods of data exchange - PDU based and signal based gateway functions. A PDU gateway routes the protocol data units (PDUs) unchanged between two networks. In this case the data carried on both networks, source and destination, are identical regarding content and length. It is also possible, that only signals, which are contained in the received PDUs from source network, are needed on the other network. In this case, the gateway does not transfer the entire PDU, but sends the individual signals to the corresponding destination network. To achieve this, a single received PDU is disassembled in signals according to the specification of the source network. Afterwards signals are grouped together congruent to configuration of the destination network and assembled to PDUs, which are send across the destination network.

The central unit of the gateway model is the module *Gateway Core* shown in Figure 3. This module contains protocol unrelated functions, e.g. buffering of received messages, routing, message segmentation and disassembling messages into signals. Based on the specification of function blocks with standardized interfaces a high adaptability of the architecture is achieved allowing the extension and substitution of single functions. Therefore it is possible to analyze the system performance of several realizations differing in their buffer strategy. In contrast to the gateway core, the protocol related gateway functions, such as protocol conversion, initialization of the communica-

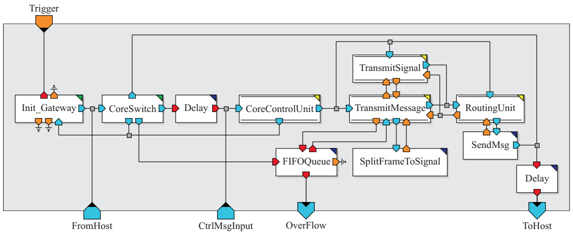


Figure 3: Module Gateway Core.

tion controller, message transmission and message reception, etc. are realized within the respective *Host* module. The host connected to the source network receives a message via the CC and converts it into an uncommitted message. Afterwards the message is sent to the gateway core. Inside the gateway core, the uncommitted message will be processed and routed to the destination host. There it is converted into the corresponding standard. Each host has a distinct identifier. This allows the connection of an unlimited number of networks to the gateway model.

The example scenario shown in Figure 4 contains two CAN and two FlexRay networks which are connected by a gateway. The system itself is an extension of an application presented in (Hedenetz and Belschner, 1998). Each network contains two or more nodes. The behavior of the used modules CCs realizes the connection to CAN and FlexRay. The analysis of the simulation results showed the correct behavior of the created model.

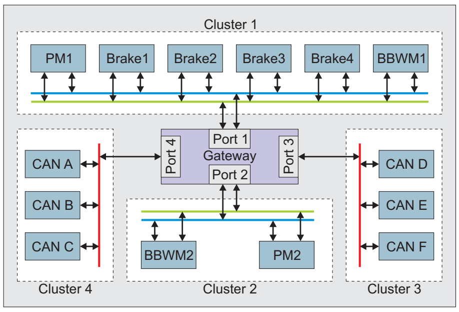


Figure 4: System used as reference for validation.

5 CONCLUSIONS

The modeling strategy and developed common gateway model presented in this paper allow the universal design and analysis of heterogeneous communication networks. Modularized components in the gateway library represent available functions of a gateway facility and allow an easy construction of a wide range of systems. The basic model concept behind the gateway model enables the analysis of communication and behavior of heterogeneous networked systems and supports the design process of such systems. In early stages of the development process the evaluation and verification of system properties can be provided prior to hardware design.

Up to now there are only few communication protocol models available. To extend the ability for system modeling it is necessary to extend the stock of models. At present this limits the systems which can be analyzed by the presented approach to CAN and FlexRay systems. To gain access to real world examples from within the automotive application domain the design of models realizing MOST (Media Oriented Systems Transport) and LIN (Local Interconnect Network) are required. Future work will also deal with the automated import of gateway routing information by using e.g. CAN Database and FIBEX (ASAM, 2007). This can be supplemented by concepts of automated model generation.

REFERENCES

ASAM (2007). *FIBEX - Field Bus Exchange Format Version 2.0.1*.

Fahmy, S. (1995). A survey of ATM switching techniques. CIS-788-95 Semester Report, Ohio State University, CIS Dep.

FlexRay Consortium (2005). *FlexRay Communications Systems - Protocol Specification Version 2.1*.

Hedenetz, B. and Belschner, R. (1998). Brake-by-wire without mechanical backup by using a ttp-communication network. page 2.

Henia, R., Hamann, A., Jersak, M., Racu, R., Richter, K., and Ernst, R. (2005). System level performance analysis - the symta/s approach. In *IEE Proceedings Computers and Digital Techniques*.

Hörner, H. (2007). Das universelle Gateway-Steuergerät. *Elektronik Automotive, Sonderausgabe AUTOSAR*, pages 33–35.

Klöckner, J., Köhler, S., and Fengler, W. (2008). Model based design of networked embedded systems. In *ICINCO-SPSMC*, pages 253–259. INSTICC Press.

MLDesign Technologies Inc. (2007). *MLDesigner Documentation, Version 2.7*. http://www.mldesigner.com/.

Müller, M. (2007). Dokumentation framebasiertes canmodell.

Salzwedel, H. (2004). Design technology development towards mission level design. In *49. Internationales Wissenschaftliches Kolloquium IWK'2004*.

Shaheen, S., Heffernan, D., and Leen, G. (2007). A gateway for time-triggered control networks. *Microprocess. Microsyst.*, 31(1):38–50.

The Ptolemy Project (2009). http://ptolemy.eecs.berkeley.edu/.

INTEGRATING REUSABLE CONCEPTS INTO A REFERENCE ARCHITECTURE DESIGN OF COMPLEX EMBEDDED SYSTEMS

Liliana Dobrica
University Politehnica of Bucharest, Faculty of Automation and Computers
Spl. Independentei 313, Bucharest, Romania
liliana@aii.pub.ro

Keywords: Embedded systems, Software architecture, Reference architecture, Patterns, Quality.

Abstract: The content of this paper addresses the issues regarding integrating reusable concepts for a quality-based design of reference architecture in the context of complexity that is specific to today's embedded control systems. The reference architecture consists of core services and is designed based on considering taxonomy of requirements and constraints, reusable control patterns and a quality-based measurement instrument.

1 INTRODUCTION

Nowadays an embedded system (ES) application represents one of the most challenging development domains. Among the requirements and constraints that have to be satisfied we can mention a higher diversity and complexity of systems and components, increased quality, standardization, fault tolerance and robustness. In the design process an ES requires introduction of the higher level abstractions that are blurring the boundaries between hardware and software design. Due to the escalating complexity level of ESs a coherent and integrated development strategy is required. It becomes a priority the creation of reference architecture (RA) and a suite of abstract components with which new developments in various application domains can be engineered with minimal effort. RA is based on a common architectural style that provides the composition of independently subsystems that meet the requirements of the various application domains. Thus different components can be created for various specific domains, while retaining the capability of component reuse across these domains.

ES complexity resides in a multitude of interdependent elements which must be organized. To handle complexity, an architectural approach helps to consider separation of concerns realized through different levels of abstraction, dynamism and aggregation levels. In the field of control, the knowledge acquired in software engineering is not really exploited, although it helps to manage complexity. Patterns and quality based approach may be used to establish a direct link between the concepts from the field of control and the software architecture concepts. They guide the analysis and synthesis of software components and they can be used to develop complex control architecture. The architecture is comprehensible as it shows the elements necessary for doing a functionality and the manner in which they interact, and it is flexible because it can be adapted to other systems of the same type in the application domain. In the context of control systems the problem is modelling and documenting software architectures reusable knowledge dedicated to control.

In this paper we propose an approach to manage complexity of complex ES based on defining sources of knowledge for RA. Building the RA is based on well known and reusable concepts from software engineering. Our contribution is in the synthesis of the most important issues that can be applied.

2 BACKGROUND

At this moment there is no general consensus about the definition of embedded terms. ESs are subject to limited memory and processing power and many ESs are also real-time systems that have strict performance constraints. Even for non-real time

ESs, developers have to take into account the timeliness, robustness, and safety of the systems. The fact that ESs are embedded, that is they cannot easily be taken out of their environment to be maintained or evolved, poses reliability requirements. Nevertheless it includes subcategories such as embedded domain, reactive domain, control/command domain, intensive data flow computation domain, best-effort services domain (Marte, 2008). Traditionally an ES represents a computer system which is integrated into another system, the embedding system. The requirements for an ES must be derived from the embedding system. There are two different areas. One is when the embedding system is a product and the other is when it is a production system. The fist one includes automotive electronics, avionics, and health care systems and the second one includes manufacturing control, chemical process control, and logistics.

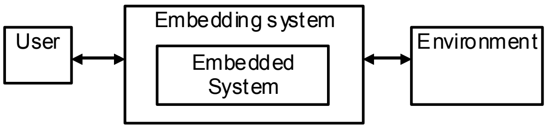


Figure 1: Traditional embedded system model.

ESs are doing control such as measuring physical data (sensing), storing data, processing sensors signals and data, influencing physical variables (actuating), monitoring, supervising, enable manual and automatic operation, etc..

In the embedded world a model driven approach is used to express the requirements in a modeling environment that automatically generates the application code. The well-known example of such an environment is the Matlab tool suite. The increase in efficiency arises from the fact that the software design and implementation phases are automated and the control engineer has not care about the implementation issues as in software engineering processes.

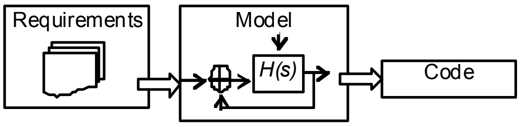


Figure2: Typical model-driven approach.

The problem for control engineering domain is that these applications tend to be multi-domain. A complete control application does not simply cover implementation of control laws. In most cases, the implementation of control laws, the specific domain of Matlab, is only a small fraction of the total control software. Most of the software normally is concerned with various functionalities and Matlab-like tools are inappropriate to cover these functionalities. A new approach is required to deal with the new requirements.

3 PROPOSED APROACH

The design of RA for complex ES is realized with core services which are abstract architectural models and depends on the quality attributes, styles and patterns and others that are shown in Figure 3.

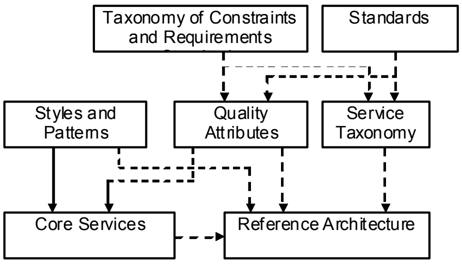


Figure 3: Reference architecture realization.

Quality attributes clarify their meaning and importance for core services. The interest of the quality attributes for the RA is how they interact and constrain each other (i.e., trade-offs) and what the user's view of quality is. The styles and patterns are the starting point for architecture development. Architectural styles and patterns are utilized to achieve qualities. The style is determined by a set of component types, the topological layout of the components, a set of semantic constraints and a set of connectors. A style defines a class of architectures and is an abstraction for a set of architectures that meet it. Design patterns are on a detailed level. They refine single components and their relationships in a particular context.

RA creates the framework from which the architecture of new ESs is developed. It provides generic core services and imposes an architectural style for constraining specific domain services in such a way that the final product is understandable, maintainable, extensible, and can be built cost-effectively. Potential reusability is highest on RA level. RA is build based on a service taxonomy. A reusable knowledge base is integrated and adapted to service engineering for ESs. The standards related to each ES domain, applicable architectural styles and patterns and existing concepts of services and components are the driving forces of ESs development. A service taxonomy defines the main categories called domains. Typical features that have been abstracted from requirements and constraints characterize services. The service taxonomy guides the developers on a certain domain and getting assistance in identifying the required supporting services and features of services.

4 DISCUSSIONS

A *taxonomy of constraints and requirements* that delimit the design space for a RA for ES is presented in figure 4. Composability refers to the way that larger systems can be composed of smaller subsystems. A system is composable with respect to a certain property if this property is not invalidated by integration. Integration of subsystems that are realized in different technologies are subject heterogeneity. Growth and scalability require if the available resources permits the integration of more subsystems, then the new ones must not disturb the correct operation of the already integrated subsystems. Integration of distributed services must adhere to well established standards.

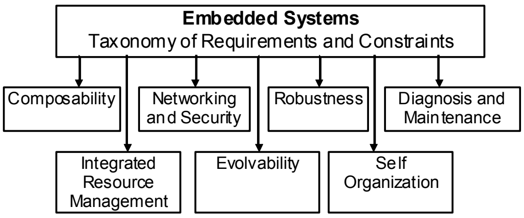

Embedded Systems
Taxonomy of Requirements and Constraints
Composability
Networking and Security
Robustness
Diagnosis and Maintenance
Integrated Resource Management
Evolvability
Self Organization

Figure 4: ESs requirements and constraints.

Networking refers to control loops to be supported at network level. Communication service reliability depends on the application parameters, and protocol standards (Ethernet, USB, CAN, Bluetooth, etc). Integrity mechanisms are required to prevent undetected modification of hardware and software by unauthorized persons or systems, meaning defence against message injections, message replay or message delay on the network. By robustness an ES must handle the increasing failure rate. Fault tolerant mechanisms are used to adapt to reliability changes of subsystems during the ES's life time. Services should be provided for error containment, membership, error detection and error masking. A generic fault-tolerance layer, design for verifiability, formal methods and specification support, software management methods for time, space, and I/O allocations should be considered, too. Diagnosis and maintenance requires a system health monitoring service and a diagnostic service to identify faulty subsystems. The diagnostic service must not interfere with the operation of the subsystems that are to be diagnosed. Predictive maintenance at the architecture level supports the identification of components that are likely to fail in the near future. Design for testability with respect to unit testing, system integration testing, manufacturing testing and assembly testing. Integrated resource management needs dynamic reconfiguration to support changing of the configurations of applications while they are executed. Evolvability is based on uncertainty with respect to application characteristics and technological capabilities. Development of products delivered in multiple variants should be considered. Implementation independence, virtual machines, legacy integration, auto-integration, and test reuse for reusable design core are included. Verification reuse defines verification patterns and environments for the subsystems at different abstraction levels. Self organizations support ubiquitous secure connectivity, mobile ad-hoc networks, and ability to adapt to user-specific behaviour.

Design and architectural patterns are important concepts in the field of software architectures to design applications by reusing generic design schemas established from successful and effective solutions. A great number of software applications are based on the same principles and their knowledge allow design efforts to be reduced considerably. Today, the main patterns are described in catalogues (Gamma et. al., 1994), (Buschmann et al., 1996). These catalogues describe the styles of organization and interaction at a higher level of abstraction, by presenting layered architectures, for example. A basic pattern for control is Strategy pattern. This separates the control from function to protect a client from various strategy services that it requires. Composite pattern is used in situations where it is necessary to treat components uniformly, regardless of whether they are primitive or composite. From behaviour perspective we can mention Chain of Responsibility pattern. Recursive control pattern (Selic, 1998) explains how to specify, then create hierarchic control architectures that are more flexible and more robust. This separates control aspects and the service providing aspects of a real time system allowing each to be defined and modified separately. The applicability of this pattern is across a wide range of levels and scopes starting from the highest system architectural level to individual components. This is useful in situations, typical in event driven real time applications where a complex software based server needs to be controlled dynamically in a non-trivial manner, where control policies may change over time. The Recursive Control is structurally related to the

Composite. It simplifies the implementation of complex systems by applying hierarchically a single structural pattern. Also it simplifies the development (and understanding) of both functional and control aspects by decoupling them from each other. It allows control or diagnosis services policies to be changed without affecting the basic functionality.

A *quality* based design requires a measurement instrument that must be defined by a taxonomy for quality attributes, which is organized with respect to three main elements: (1) The priority in a quality attributes list. The presence of this element in the taxonomy is necessary, due to the costs required by an analysis method at the architectural level. (2) Architecture views which are relevant for that quality attribute; (3) Appropriate methods to be applied for quality attribute analysis.

Quality attributes may be classified in essential, very desirable, desirable, don't care and forbidden. The priorities are established based on the experts' knowledge and the stakeholders' objectives. Quality function deployment (Reed, 1993) is a suitable technique for showing the relational strengths from objectives of stakeholders and architectural to quality attributes. These priorities are important for the evaluation process, which considers an analysis method for each quality attribute. At this moment various architecture analysis methods, such as scenario-based architecture analysis (SAAM) (Kazman et al. 1994), architecture tradeoff analysis (ATAM) (Kazman et al, 1998), architecture level prediction of software maintenance (ALPSM) (Bengston, 2004), or reliability analysis using failure scenario (SARAH) exist. Methods are distinguished by the evaluation techniques, the number of quality attributes and their interaction for tradeoff decisions, the stakeholders' involvement, and how detailed the architecture design is at the moment the analysis (Dobrica and Niemela, 2002).

The measurement instrument is applied to the RA during analysis. The quality attribute with the first priority in a list is first analyzed with respect to the appropriate architecture view and the appropriate method. Then the next quality attribute from the list is analyzed in isolation and then considering the interaction with the first one for finding sensitivity points and tradeoffs on the services included in the RA. The process is repeated for all the attributes in the list. In order to decide on RA core services, this procedure could also be improved and refined. In this case special attention should be paid to the collections of services in the architecture which are critical for achieving a particular quality attribute, or architectural elements to which multiple quality attributes are sensitive. A deeper level of analysis could influence the decision on the addition of new services to the RA.

5 CONCLUSIONS

This paper has proposed an approach for a RA development for complex ES application domains based on a knowledge of reusable concepts from software engineering at architectural level. The approach has an immense potential to improve embedded control systems development as well as reduce time and costs in stages such as architecture design and analysis. However, for this approach's success it is necessary to create a cooperation culture among embedded control system developers. Future research work is needed to develop systematic ways of bridging these reusable concepts to a RA, reducing in this way the cognitive complexity.

ACKNOWLEDGEMENTS

This work is supported by the Romanian research grant CNCSIS IDEI no. 1238/2008.

REFERENCES

MARTE, 2008, Modeling and Analysis of Real Time and Embedded, www.omg.org.

Buschmann F., R. Meunier, and H. Rohnert, 1996, *Pattern-Oriented Software Architecture: A System of Patterns*, John Wiley and Sons.

Gamma E., R. Helm, R. Johnson, and J. Vlissides, 1994, *Design Patterns: Elements of Reusable Object-Oriented Software*, Addison Wesley.

Selic B., 1998, Recursive control, in: R. Martin, et al. (Eds.), *Patterns Languages of Program Design*, Addison-Wesley, pp. 147–162.

Kazman R., L. Bass, G. Abowd, M. Webb,1994, SAAM: A method for analyzing the properties of Software Architectures, *Procs of the ICSE*, 81-90.

Kazman R., M. Klein, M. Barbacci, H. Lipson, T. Longstaff, S. J. Carrière, 1998, The Architecture Tradeoff Analysis Method, *Procs. of the ICECCS.*

Reed B.M., D.A. Jacobs, 1993, *Quality Function Deployment For Large Space Systems*, National Aeronautics and Space Administration.

Bengston PO, 2004, Architecture Level Prediction of Software Maintenence, *Procs of the ICSR5*.

Dobrica L. and Niemelä E., 2002, A survey on software architecture analysis methods, *IEEE Transactions on Software Engineering*, 28(7), 638-653.

PERFORMANCE EVALUATION OF A MODIFIED SUBBAND NOISE CANCELLATION SYSTEM IN A NOISY ENVIRONMENT

Ali O. Abid Noor, Salina Abdul Samad and Aini Hussain
Department of Electrical, Electronic and System Engineering, Faculty of Engineering and Built Environment
University Kebangsaan Malaysia – UKM, Malaysia
{ali511, salina, aini}@vlsi.eng.ukm.my

Keywords: Noise Cancellation, Adaptive Filtering, Filter Banks.

Abstract: This paper presents a subband noise canceller with reduced residual noise. The canceller is developed by modifying and optimizing an existing multirate filter bank that is used to improve the performance of a conventional full-band adaptive filtering. The proposed system is aimed to overcome problems of slow asymptotic convergence and high residual noise incorporating with the use of oversampled filter banks for acoustic noise cancellation applications. Analysis and synthesis filters are optimized for minimum amplitude distortion. The proposed scheme offers a simplified structure that without employing cross adaptive filters or stop band filters reduces the effect of coloured components near the band edges in the frequency response of the analysis filters. Issues of increasing convergence speed and decreasing the residual noise at the system output are addressed. Performance under white and coloured environments is evaluated in terms of mean square error MSE performance. Fast initial convergence was obtained with this modification. Also a decrease in the amount of residual noise by approximately 10dB compared to an equivalent subband model without modification was reachable under actual speech and background noise.

1 INTRODUCTION

Subband adaptive filtering using multirate filter banks has been proposed in recent years to speed up the convergence rate of the least mean square LMS adaptive filter and to reduce the computational expenses in acoustic environments (Petraglia and Batalheiro, 2008). In this approach, multirate filter banks are used to split the input signal into a number of frequency bands, each serving as an input to a separate adaptive filter. The subband decomposition greatly reduces the update rate of the adaptive filters, resulting in a much lower computational complexity. Furthermore, subband signals are often downsampled in a subband adaptive filter system, this leads to a whitening effect of the input signals and hence an improved convergence behavior.

In critically sampled filter banks, where the number of subbands equals to the downsampling factor, the presence of aliasing distortions requires the use of adaptive cross filters between subbands (Petraglia et al., 2000). However systems with cross adaptive filters generally converge slowly and have high computational cost, while gap filter banks produce spectral holes which in turn lead to significant signal distortion. Problems incorporating with subband splitting have been treated in literature regarding issues of increasing convergence rate (Lee and Gan, 2004), lowering computational complexity (Schüldt et al., 2000) and reducing input/output delay (Ohno and Sakai, 1999).

Oversampled filter banks has been proposed as the most appropriate solution to avoid aliasing distortion associated with the use of critically sampled filter banks (Cedric et al., 2006) . However this solution implies higher computational requirements than critically sampled one. In addition, it has been demonstrated in literature that oversampled filter banks themselves color the input signal, which leads to under modelling (Sheikhzaheh et al., 2003). These problems can be traced back to the fact that oversampled subband input will likely generate an ill-conditioned correlation matrix (Deleon and Etter, 1995). In this case, the small eigenvalues are generated by the roll off of the subband input power spectrum. A pre-emphasis filter for each subband is suggested by (Tam et al., 2002) as a remedy for this slow asymptotic convergence. An alternative approach to remove the band edge components might be the use of a bandstop filters. But this has the undesirable effect of introducing spectral gaps in the reconstructed full band signal. Furthermore it was proved by (Sheikhzaheh et al., 2003) that the introduction of pre-emphasis filters has no considerable effects on the convergence behaviour of a subband noise canceller.

In this paper an alternative procedure is adopted to improve the performance of a noise cancellation system aimed to remove background noise from speech signals. Different prototype filters are used in the analysis and synthesis filter banks. The analysis prototype filter is modified so that the coloured components near the band edges are removed by synthesis filtering, and then the analysis/synthesis filter bank is optimized in the input/output relationship to achieve minimum amplitude distortion. Compared to literature designs (Deleon and Etter, 1995) and (Cedric et al., 2006), this paper bears three differences: first, different methods used for the design of analysis and synthesis filter banks, second, optimization is performed to reduce amplitude distortion with negligible aliasing error due to the use of highly oversampled filter banks and third, the resulting design is implemented efficiently for the removal of background noise from speech signals.

The proposed modified oversampled subband noise canceller offers a simplified structure that without employing cross-filters or gap filter banks decreases the residual noise at the system output. Issues of increasing convergence rate and reducing residual noise on steady state are addressed. Performance under white and coloured environments is evaluated in terms of mean square error MSE convergence. Comparison is made with a conventional full band scheme as well as with a similar system with no modification. The paper is organized as follows: in addition to this section, section 2, formulates the subband noise cancellation problem, section 3 describes the optimum analysis/synthesis filters design, section 4 presents simulation results with discusses the main aspects of the results and section 5 warps up the paper with concluding remarks.

2 SUBBAND NOISE CANCELLER

The original noise cancellation model described in (Sayed, 2008) is extended to subband configuration by the insertion of analysis/synthesis filter banks in signal paths, as depicted by Figure 1.

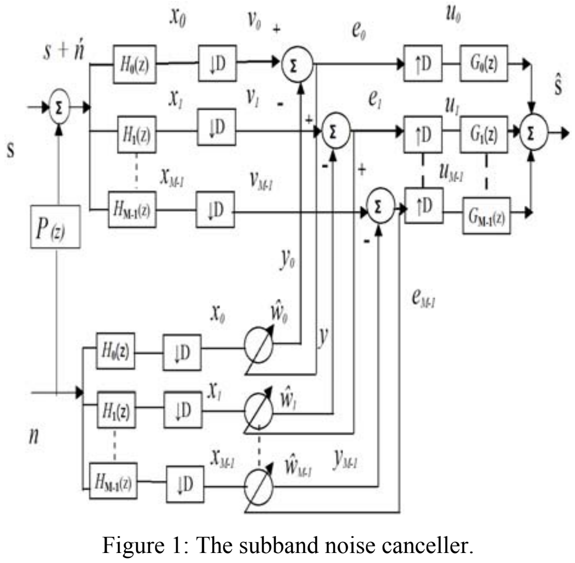


Figure 1: The subband noise canceller.

Assigning uppercase letters for z-transform representation of variables and processors, the noisy speech and the background noise are split into subbands by two sets of analysis filters. The subband analysis filters is expressed in z-domain as

$$H_k(z) = \sum_{m=0}^{L-1} h_k(m) z^{-m} \tag{1}$$

for *k=0,1,2,....M-1.*

Where *k* is the decomposition index, *h(m)* is the impulse response of a finite impulse response filter FIR , *m* is a time index, *M* is the number of subbands and *L* is the filter length. Now, consider the adaptation process in each individual branch according to Figure 1, and let us define *e(m)* as the error signal, *y(m)* is the output of the adaptive filter calculated at the downsampled rate, $\hat{\mathbf{w}}(m)$ is the filter coefficient vector at *mth* iteration, μ is the adaptation step-size factor , α is proportional to the inverse of the power input to the adaptive filter, and *m* is a time index, then we have

$$y_k(m) = \hat{\mathbf{w}}^T{}_k(m)\mathbf{x}_k(m) \tag{2}$$

$$e_k(m) = v_k(m) - y_k(m) \tag{3}$$

$$\hat{\mathbf{w}}_k(m+1) = \hat{\mathbf{w}}_k(m) + \mu_k \alpha_k e_k(m)\mathbf{x}_k(m) \tag{4}$$

Relation (4) represents the subband normalized LMS update of the branch adaptive filters. In z-domain, relation (3) can be expressed as

$$E_k(z) = V_k(z) - Y_k(z) \tag{5}$$

Where $V_k(z)$ represents the downsample subband noisy signal, and $Y_k(z)$ is the output of the subband adaptive filter. The aim of the adaptive process is to surpress the noisy component in $V_k(z)$ by equating it to $Y_k(z)$ leaving the subband input $S_k(z)$ undistorted. Each subband error signal is then interpolated by upsampling and synthesis filtering. The final output can be expressed as

$$\hat{\mathrm{S}}_k(z) = \sum_{k=0}^{M-1} G_K(z)U_k(z) \qquad (6)$$

Where $U_k(z)$ represents the upsampled version of the subband error signals $E_k(z)$.We assume a highly oversampled filter bank, say two fold , the aliasing components in (6) can be considered to be very small and therefore neglected, hence the final system equation can be represented as

$$\hat{\mathrm{S}}_k(z) = S(z)\sum_{k=0}^{M-1} G_K(z)H_k(z) \qquad (7)$$

If we constrain the analysis and synthesis filters $H_k(z)$ and $G_k(z)$ respectively to be linear phase finite impulse response FIR filters, then the term $\sum_{k=0}^{M-1} G_K(z)H_k(z)$ in equation (7) describes the amplitude distortion function of the system.

The branch filters $H_K(z)$, and $G_K(z)$, can be derived from single prototype filters $H_0(z)$, $G_0(z)$, according to $H_k(z) = H_0(zW_M^k)$ and $G_k(z) = G_0(zW_M^k)$, where $W_M = e^{-j2\pi/M}$.

This way, a uniform discrete Fourier transform DFT filter banks are created. Let *A(z)* be the distortion function ,in frequency domain, *A(z)* can be represented as

$$A(e^{j\omega}) = \sum_{k=0}^{M-1} H_k(e^{j\omega})G_k(e^{j\omega}) \qquad (8)$$

The objective is to find prototype filters $H_0(e^{j\omega})$ and $G_0(e^{j\omega})$ to minimize $A_d(z)$ according to

$$\mathrm{A_d} = (1 - \left|A(e^{j\omega})\right|) \qquad (9)$$

Ideally $\mathrm{A_d}$ should be zero i.e. a perfect reconstruction filter bank. Relaxing the perfect reconstruction property, we can tolerate small amplitude distortion; and then we can have frequency selective filters in a near perfect reconstruction NPR filter bank.

3 ANALYSIS/SYNTHESIS PROTOTYPE FILTER DESIGN

The aim of the prototype filter design is to build a complementary analysis and syntheses filter banks, so that the reconstructed output signal has a very low distortion. Different prototype filters for analysis and synthesis filter banks are designed. The analysis and synthesis prototype filters are selected as in Figure 2. The cut off frequency f_c of the analysis prototype filter is given by

$$f_c = (f_{pa} + f_{sa})/2 \qquad (10)$$

provided that $f_{pa} > f_{ss}$.

Where f_{pa} is the end of the passband of the analysis filter, f_{sa} is the beginning of the stopband of the analysis filter; f_{ss} is the beginning of the stopband of the synthesis filter.

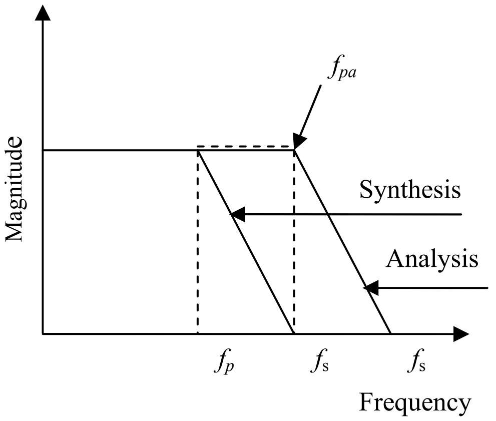


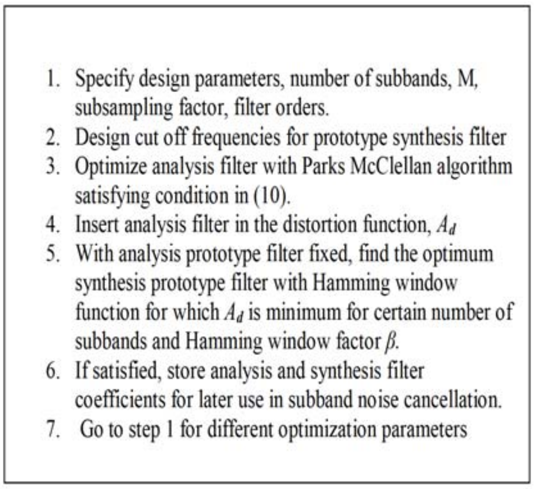
Figure 2: Analysis/Synthesis filters design.

The 3dB down of the prototype synthesis filter (normalized) (*1 /2 M*) is determined by the number of subbands *M*, whereas the analysis filter bandwidth is larger and only limited by the decimation factor *D* according to

$$D \le \frac{2}{f_{pa} + f_{sa}} \qquad (11)$$

where *D* is the largest integer less than or equal to the right hand side term of (11), f_{pa} and f_{pa} are normalized to the sampling frequency. Values for *D* such that the ratio M/D =2 i.e. two fold over sampled found to be the best in terms of alias cancellation.

The design algorithm starts by designing cut off frequencies for an arbitrary analysis and synthesis prototype filters. The analysis prototype filter is optimized using Parks McClellan algorithm to meet requirements in (10). With analysis prototype filter fixed, the synthesis prototype filter is optimized by minimizing the distortion function given by (9) over the frequency grid in the band [0-π].Figure 3 depicts steps the design procedure and Figure 4 shows a reconstruction error comparison between the modified oversampled filter bank and an equivalent conventional oversampled filter bank.



1. Specify design parameters, number of subbands, M, subsampling factor, filter orders.
2. Design cut off frequencies for prototype synthesis filter
3. Optimize analysis filter with Parks McClellan algorithm satisfying condition in (10).
4. Insert analysis filter in the distortion function, A_d
5. With analysis prototype filter fixed, find the optimum synthesis prototype filter with Hamming window function for which A_d is minimum for certain number of subbands and Hamming window factor β.
6. If satisfied, store analysis and synthesis filter coefficients for later use in subband noise cancellation.
7. Go to step 1 for different optimization parameters

Figure 3: Steps of design algorithm.

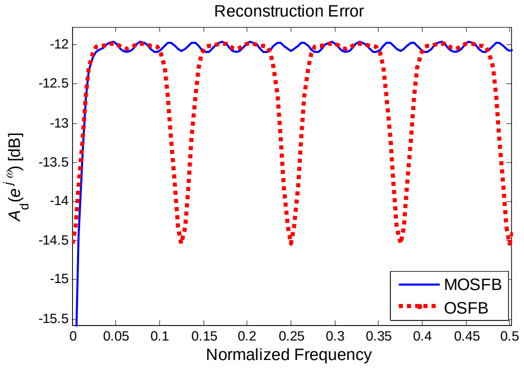


Figure 4: Reconstruction error comparison of the modified filter bank MOSFB and an equivalent conventional oversampled filter bank OSFB.

Analysis and synthesis filter banks can be implemented efficiently using polyphase representation of a single prototype filter followed by fast Fourier transforms FFT.

4 RESULTS AND DISCUSSION

The noise path used in these tests is an approximation of a small room impulse response modelled by a FIR processor of 256 taps. The number of subbands $M=8,$ downsampling factor $D=4$, prototype filters order is 128.To measure the convergence of the subband noise canceller, a variable frequency sinusoid was corrupted with white Gaussian noise that was passed through a transfer function representing the acoustic path. The corrupted signal was then applied to the primary input of the noise canceller; regarding zero mean, white Gaussian noise is applied to the reference input. A subband power normalized version of the LMS algorithm is used for adaptation. Mean square error MSE convergence is used as a measure of performance. Plots of MSE were produced and smoothed with a suitable moving average filter. A comparison is made with a conventional fullband system as well as with an oversampled system without modification. The unmodified oversampled subband noise canceller is denoted with OSSNC and the modified system is denoted with MOSSNC. Results are depicted in Figures 5 and 6.

To test the behaviour under real environmental conditions, a speech signal is then applied to the primary input of the proposed noise canceller. The speech is a Malay utterance “Kosong, Satu, Dua, Tiga” spoken by a woman, sampled at 16 kHz. Different types of background interference were used to corrupt the aforementioned speech. MSE plots are produced for two cases: Figure 7 for the case of machinery noise as background interference and in Figure 8 for the case of a cocktail party i.e. disturbance by another speech.

Apart from the fast initial convergence, it is clear from Figure 5, that the mean square error (MSE) plot of the oversampled subband system OSSNC levels off quickly before the MSE plot of the fullband system. This is obviously due to the inability to properly model the presence of coloured components near the band edges of the filter bank. During initial convergence the subband system performs better than the fullband system but is less effective afterward. On the other hand, the MSE convergence of the modified system MOSNC outperforms that of the fullband system during initial convergence and exhibits comparable steady state performance as shown by Figure 6. It is obvious that while the MSE of the fullband system converging in slow asymptotic way, the MOSNC system reaches a steady in 2000 iterations. The fullband system needs more than 6000 iterations to reach the same noise cancellation level. The main difference between Figure 5 and Figure 6 is in the amount of residual noise which has been g reduced with the MOSSNC. Results obtained for actual speech and background noise (Figures 7 and 8) prove that the fullband system cannot model properly with coloured noise as the input to the adaptive filters, and the residual error can be sever when the environment noise is highly coloured. Tests performed in this part of the experiment proved that the MOSSNC does have improved performance compared to OSSNC and the full-band model. Depending on the type of the interference, the improvement in reducing resedual noise n varies from 15-20 dB better than full-band case.

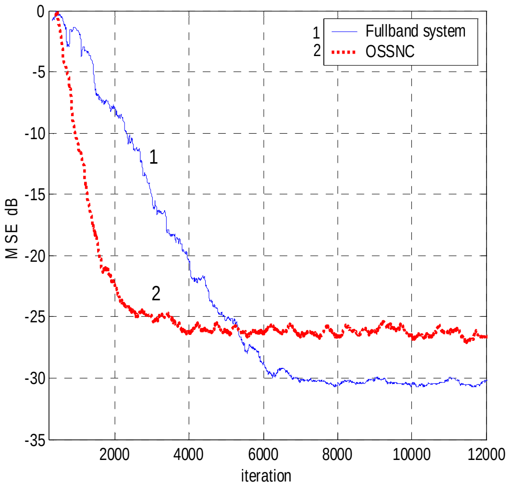


Figure 5: MSE convergence behaviour of OSSNC under white Gaussian noise.

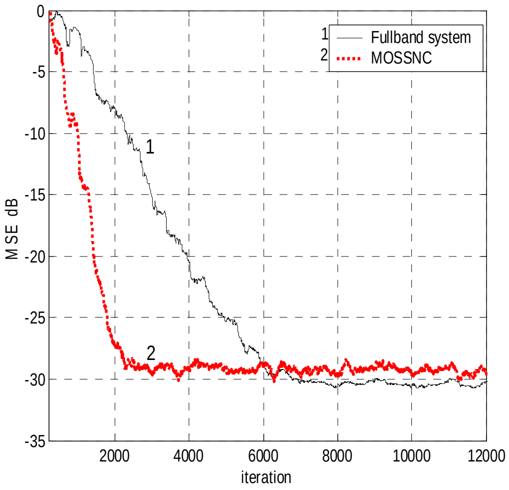


Figure 6: MSE convergence behaviour of MOSSNC under white Gaussian noise.

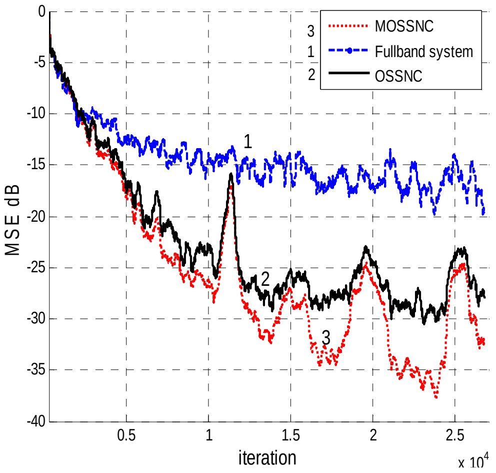


Figure 7: Performance comparison under actual speech and machinery noise.

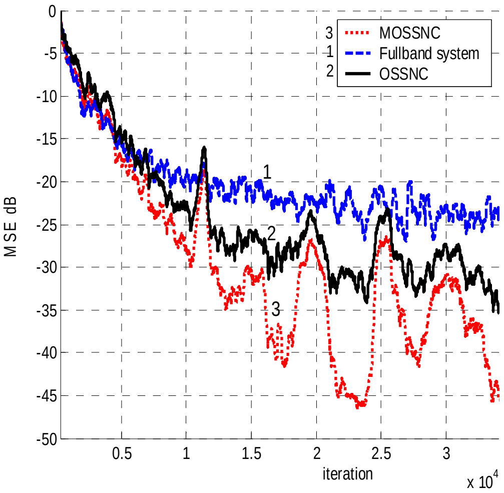


Figure 8: Performance comparison under actual speech disturbed by another speech.

5 CONCLUSIONS

In this work, an oversampled subband noise canceller with modified filter bank is developed to overcome the problem of slowly converging components associated with the usual oversampled subband scheme. An efficient optimized DFT filter bank is used in the canceller. The analysis filter bank is modified to remove slowly converging components near band edges, while the synthesis filter bank is optimized to minimize input/output distortion.The modified system has shown improved performance compared to a similar scheme with conventional oversampled filter bank. The convergence behaviour under white and coloured environments is greatly improved. The amount of residual noise is reduced by 15-10dB under actual speech and background noise. The next logical step is to realize this system on a suitable DSP processor such as the Texas C6000 to prove the validity of the method for noise cancellation. Also, other types of filter banks and transforms can be investigated and used for the same purpose.

REFERENCES

Cedric K. F., Grbic, Y. N., Nordholm S., Teo, K. L., 2006. A hybrid Method for the Design of Oversampled Uniform DFT Filter Banks. *Elsevier Signal Processing 86 (2006)*, pp1355-1364. www.elsevier.com/locate/sigpro

Deleon, P. Etter, D.,1995. Experimental Results with increased Band Width Analysis Filters in Oversampled, Subband Acoustic Echo Cancellers. *IEEE Signal Processing Letters* ,Vol. 2 No.1, January 1995, pp.1-3.

Lee, K. A., Gan, W. S., 2004. Improving Convergence of the NLMS Algorithm Using Constrained Subband Updates. *IEEE Signal Processing Letters*, Vol. 11, No. 9, Sebtember 2004, pp736-239.

Ohno, S., Sakai, H.,1999. On Delayless Subband Adaptive Filtering by Subband/Fullband Transforms. *IEEE Signal Processing Letters*, Vol. 6, No. 9, Sebtember 1999,pp. 236-239.

Petraglia, M. R., Batalheiro, P., 2008. Non-uniform Subband Adaptive Filtering With Critical Sampling, *IEEE Transactions on Signal Processing*, Vol56, No. 2, February 2008. pp565-575.

Petraglia, M. R., Rogerio G.A., Diniz, P. S. R, 2000. New Structures for Adaptive Filtering in Subbands with Critical Sampling. *IEEE Transactions on Signal Processing,* Vol. 48, No. 12, December 2000, pp3316-3323.

Sayed, H.A., 200. *Adaptive Filters*, John Wiley NJ, 2008.

Schüldt, C., Lindstrom, F., Claesson I., 2008. A Low-Complexity Delayless Selective Subband Adaptive Filtering Algorithm. *IEEE Transactions on Signal Processing*, Vol56, No. 12, December 2008, pp 5840-5850.

Sheikhzaheh, H., Abutalebi, H. , Brennan, R. L., Sollazzo, J., 2003. Performance Limitation of a New Subband Adaptive System for Noise and Echo Reduction. *Proceedings of IEEE Int. Conf. on Electronics, Circuits and Systems ICECS,* December 14-17, Volume:2, 2003, Sharjah UAE, pp 459- 462.

Tam, K., Sheikhzadeh, H., Schneider, T., 2002. Highly Oversampled Subband Adaptive Filters for Noise Cancellation on a Low-Resource DSP System. *In Proceedings. of7th Int. conf. on spoken language. Processing (ICSLP)*, Denver, Co September 2002.

MAPPING DEVELOPMENT OF MES FUNCTIONALITIES

Vladimír Modrák
Faculty of Manufacturing Technologies
Technical University of Košice, Bayerova 1, Slovakia
vladimir.modrak@tuke.sk

Ján Manduľák
LPH Vranov N/T, S.R.O., Pod dolami 838, Slovakia
jmandulak@lph.sk

Keywords: Manufacturing management, Information system, Interface, functionality.

Abstract: This paper presents a view on MES and ERP functional areas in a hierarchy of enterprise information and control systems. It starts with a background on ERP and MES Evolution. The work is based on the exploration of MES and ERP functionalities development. Consecutively, aspects of ERP and MES integration are treated. In the final section an impact of RFID technology on a validity data stored in MES obtaining from a tracing of material flow in production processes is analyzed.

1 INTRODUCTION

In the present manufacturing paradigm, manufacturing execution systems (MESs) play a significant role. Offered software solutions simultaneously close the gap between Enterprise Resource Planning (ERP) systems and production equipment control or SCADA (Supervisory Control And Data Acquisition) applications. Current ERP systems contain usually modules for material management, accounting, human resource management and all other functions that support business operations. In the past years, the role of ERP has been extended to cross-organizational coordination. Nowadays, as optimization of production activities is increasingly topical, a cooperation of ERP and MES becomes a serious concern of manufacturing managers. The paper is structured as follows. Firstly, a brief view on MES Evolution is presented. Then, MES functionalities are partially analyzed and a general functionality model is described. After that, technical aspects of ERP and MES integration are treated. Finally, decisive factors that influence the further development of manufacturing execution systems are discussed.

2 VIEW ON MES EVOLUTION

As ERP systems by nature are not suitable for controlling day to day shop floor operations, for this purpose a new type of industrial software with acronym MES has emerged during nineties (Choi and Kim, 2002). There is a more interpretation of MES depending on different manufacturing conditions, but the common characteristic to all is that an MES aims to provide an interface between an ERP system and shop floor controllers by supporting various ‘execution’ activities such as scheduling, order release, quality control, and data acquisition (MESA #6, 1997). In a context of the MES development and deployment it is important to point out that Manufacturing Execution Systems were originally designed to provide first-line supervision management with a visibility tool to manage work orders and workstation assignments. Consecutively, MES expanded into the indispensable link between the full range of enterprise stakeholders and the real-time events occurring in production and logistics processes across the extended value chain (McClellan, 2004).

The phenomena of globalization forces manufacturers to continuously improve their performance. In this context, manufacturing and operational excellence has become the key theme for the manufacturing companies. To improve their performance, most manufacturers apply methods and techniques which are focused on the elimination of non-value adding activities. Information systems can by supported in such programs or they can provide a complementary way of improving performance by increasing visibility on plant performance. Accordingly, cooperation of ERP and Manufacturing Execution Systems (MES) becomes a serious concern of manufacturing managers. In that sense, from MES applications is expected to support real-time production control as well as data collection and reporting to facilitate information operability in a company.

3 MES FUNCTIONALITIES

The A concept of Manufacturing Execution Systems is one of several major information systems types aimed at manufacturing companies. MES can be in simple way also defined as a toll for manufacturing management. The functions of an MES range from operation scheduling to production genealogy, to labour and maintenance management, to performance analysis, and to other function in between. There are several general models of typical MES functions that are principally divided into core and support functions (see more in Modrák, 2005). The core functions deal primarily with actual management of the work orders and the manufacturing resources. Other functional capabilities of MES may be required to cover support aspects of the manufacturing operations.

MESA International presents another approach to MES functionalities that is more or less based on the assumption of profitability to begin to deal with wider model of basic elements to ensure incorporating all-important functions into MES (MESA #2,1997).

A point of debate about MES functionalities also is connected with different types of manufacturing.

Understandably, from automation point of view a discrete manufacturing presents much more complicate concept comprising of various technologies that are used to integrate manufacturing system to one another. As the aim of this work is to generalize MES functionalities it is also reasonable to model of hierarchical levels and functions in a common manufacturing company. A hierarchical structure of main companies' functions in this case can be represented by four levels (see figure 1).

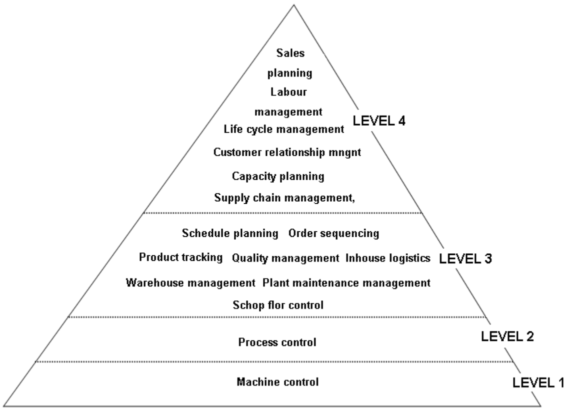


Figure 1: Functional levels in a manufacturing company.

Model of such structured company functions is often divided into three levels that are the company management, the production management, and the production control (Gunther et al, 2008). In this relation, functional areas of MES and ERP might not be considered as closed structure, because it was recognized that functions can run in the classic ERP environment as well as in the MES environment. Accordingly, under specific circumstance they may overlap of both systems. Based on this assumptions the following structure of MES and ERP functions depicted in Figure 2 is mapped.

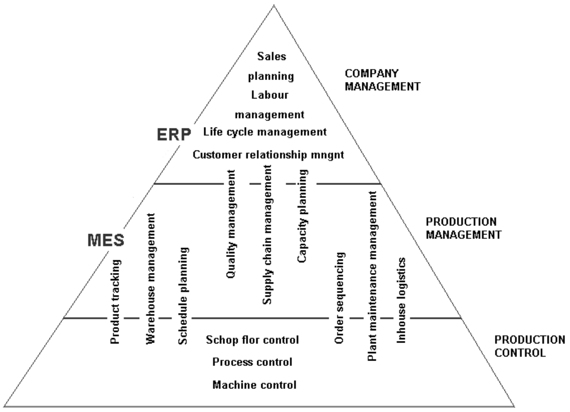


Figure 2: Intersections of MES and ERP functional areas (adapted from Gunther et al, 2008).

Obviously, the scope of operations or functions depends on number of subsystems, but the key functions remain unchanging in their essence. Because, there are no reference MES models that can be used for general manufacturing environments, overcoming of this aspect leads through the presentations of sample solutions by types of environment and other criterions. As example can be used approach to modeling three different management systems for maintenance, quality and production (Brandl, 2002) based on the S95 standard of ISA (ANS/ISA, 2000).

4 CHALLENGES OF ERP AND MES INTEGRATION

Manufacturing execution systems besides their typical functions were developed and used also as the interface between ERP and process control, since it was generally recognized that ERP systems weren't scalable. The seamless connections often required skilled coding to connect to ERP and process control systems (Siemens Energy & Automation, Inc., 2006). Today, the availability of Web-based XML communications successfully bridges the gaps between MES and ERP systems. Built on XML, the B2MML (business-to-manufacturing markup language) standard specifies accepted definitions and data formats for information exchange between systems, and facilitates information flow and updates between ERP and manufacturing execution systems. It also instigated redefinition the role of the MES. The ISA SP-95 model (see Figure 3) breaks down business to plant floor operations into four levels.

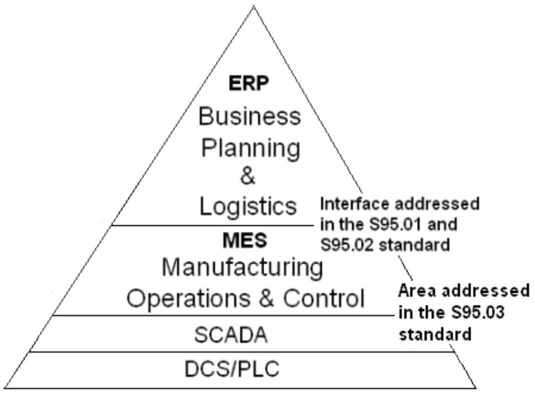

ERP
Business Planning & Logistics
Interface addressed in the S95.01 and S95.02 standard
MES
Manufacturing Operations & Control
Area addressed in the S95.03 standard
SCADA
DCS/PLC

Figure 3: ISA SP 95 control hierarchy.

Levels 1 and 2 include process control zone. MES layer consists of managerial and control functions depending on different types of manufacturing. Level 4 corresponds to the business planning and logistics.

The goal of ISA-95 standard was to reduce the risk, cost and errors associated with implementing interface between ERP and MES. The ISA-95 "Enterprise - Control System Integration" is a multi-part series of ANSI/ISA standards that define the activity models and interfaces between manufacturing functions and other enterprise functions. Parts 1 (Models and Terminology), parts 2 (Objects Attributes) and part 5 (Business to Manufacturing Transactions) define the exchange of production data between business and plant systems. B2MM provides a schema implementation of the ANSI/ISA-95 and represents an independent technology implementation of this standard. B2MML has been developed by The World Batch Forum (WBF) and adopted by players such as SAP and Wonderware. Coupled together, B2MML and ISA-95 permit designers to bridge ERP and MES systems by using B2MML XML vocabulary.

Mentioned and other ISA standards significantly facilitate the implementation of integrated manufacturing systems. It is aimed to integrate ERP systems with control systems like DCS and SCADA. To support batch control level optimization, the standard S88.01 (ANSI/ISA, 1995) has been developed. It provides standard models and terminology for the design and operation of batch control systems. At the control level the key attribute is integration of all process information into one place. For this purpose are ordinarily used both a programmable logic controllers and SCADA software.

5 CHALLENGES OF ERP AND MES INTEGRATION

An effectiveness of exploitation of new manufacturing technologies depends on the way how successfully will be synchronized newly obtained data from a production control layer into MES/ERP systems. This challenge escalates as the RFID applications are increasing to a large number of products and facilities and as they include integration in broader Supply Chain Management systems. According to Williams (2005), the opportunities enabled by RFID are expected apart from other effects in simplification of business processes. Many manufacturing organizations have processes where a product, asset, document or even a person is "touched" by many different people at different times. It causes limited view of information that can introduce inefficiencies in the overall process when information about other steps is needed to execute the current step. Accordingly, common MES/ERP systems can not have an access to detailed information and they have no idea of what is really happening to material flow on the shop floor. Mentioned drawback leads to insufficient coordination between material and accompanying information flows and so-called bull-whip-effect. When all data that information systems operate with are "fed" to them by intermediary subject, information on material flow is time dependent so it is already outdated when inserted into the information system by human operator. Until the next synchronization information become more and more outdated. Reducing the bull whip effect by means of RFID system improves the efficiency of execution/information systems not only within the site but also across the supply chain. The results of our experiments presented earlier (Modrák and Moskvič, 2007) showed that application of RFID technology for tracking and traceability of material flow will impact the whole performance of information systems in terms of information validity and practically eliminate time dependence of amount and quality of information available for ERP/MES systems.

6 CONCLUSIONS

As it is conceded that production planning activities have become more complex and therefore need to be in principle optimized. Manufacturing Execution Systems, which are positioned between the Enterprise Resource Planning and control systems levels, have significant potential to be effectively used to optimize business processes on the shop floor. Besides that fact, MES are being viewed as critical in getting the most value out of existing investments in automation. A frequent interest of manufacturers concerns a balanced scale of MES functionalities. As mentioned earlier, it depends on more factors. For instance, when an existing ERP system contains factory floor control functionality, then functionality model of MES has only supplement character. Thus, a scope of MES functionality is evidently influenced by changes in using automated identification (AID) technologies, because they have positive impact on the plant floor optimization. Therefore, mass use of RFID technology can bring significant rationalizations in the manufacturing automation in the near future. This tendency was indirectly confirmed by such IT players as Oracle, SAP, Microsoft and IBM, as they all have accelerated efforts to meet the RFID challenge (Rockwell Automation, 2004). In this sense, rules concerning manufacturing execution such as control, scheduling, routing, tracking, and monitoring might all be modified responding to RFID challenges.

REFERENCES

ANSI/ISA S88.01, 1995. Batch Control Part 1: Models and Terminology, *International Society for Measurement and Control*, RTP North Carolina, USA.

ANS/ISA-95.00.03, 2000. Enterprise Control System Integration Part 3: Models of Manufacturing Operations, Draft 7, *International Society for Measurement and Control*, RTP North Carolina, USA.

Brandl, D., 2002. Making Sense of the MES at the MES Layer" In *ISA, Technical Conference*, Chicago IL, October.

Choi, B.K., and Kim, B.H., 2002. MES architecture for FMS compatible to ERP, *Int. Journ. of Computer Integrated Manufacturing,* 15, (3), pp. 274-284.

Gunther, O., Kletti, W., Kubach, U., 2008. *RFID in Manufacturing*, Springer Berlin and Heidelberg GmbH.

McClellan, M., 2004. Execution Systems: The Heart of Intelligent Manufacturing, *Intelligent Enterprise*, Jun 12.

MESA #2, 1997. MES functionalities and MRP to MES Data Flow Possibilities. *White Paper 2,* Update and Revised March 1977, Manufacturing Execution Systems Association, Pittsburgh, P.A.

MESA #6, 1997. MES explained: a high level vision. *White Paper 6*, Manufacturing Execution Systems Association, Pitsburgh, P.A.

Modrák, V., 2005. Functionalities and Position of Manufacturing Execution Systems. In *M. Koshrow-Pour (Ed.) Encyclopedia of Information Science and Technology. First Edition*, Idea Group Reference, Hershey, USA.

Moskvič, V., and Modrák, V., 2007. RFID In Automotive Supply Chain Processes -There is a Case. In *T. Sobh, K. Elleithy, A. Mahmood, M. Karim (Eds.) Innovative Algorithms and Techniques in Automation, Industrial Electronics and Telecommunications*. Springer, New York, USA,

Rockwell Automation 2004. "RFID in manufacturing", white paper. Retrieved from www.rockwel automation.com/solutions/rfid/get/rfidwhite.pdf

Siemens Energy & Automation, Inc., 2006. Why integrate MES and ERP? Because you can't afford not to. *Siemens Whitepaper*, 1-8.

Williams, D.H., 2005. Beyond the Supply Chain: The Impact of RFID on Business Operations and IT Infrastructure. *Planetpal whitepaper*. Retrieved from http://www. planetpal.net /En/infos/art26_05_05.shtm

AUTHOR INDEX

AUTHOR INDEX (CONT.)

Proceedings of ICINCO 2009
6th International Conference on Informatics in Control, Automation and Robotics - Volume 3
ISBN: 978-989-674-001-6
http://www.icinco.org