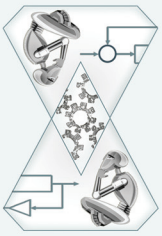


# ICINCO 2008

## FIFTH INTERNATIONAL CONFERENCE ON INFORMATICS IN CONTROL, AUTOMATION AND ROBOTICS

# Proceedings

Signal Processing, Systems Modeling and Control

FUNCHAL, MADEIRA - PORTUGAL · MAY 11 - 15, 2008

CO-ORGANIZED BY

INSTICC

UNIVERSIDADE da MADEIRA

IN COOPERATION WITH

aaai

CO-SPONSORED BY

IEEE

IEEE Systems, Man, and Cybernetics (SMC) Society

IFAC

INSTICC PRESS

# ICINCO 2008

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

Volume SPSMC

Funchal, Madeira, Portugal

May 11 – 15, 2008

Co-organized by
**INSTICC – Institute for Systems and Technologies of Information, Control and Communication**
**and**
**UMa – Universidade da Madeira**

Co-sponsored
**IEEE SMC – IEEE Systems, Man and Cybernetics Society**
**and**
**IFAC – International Federation of Automatic Control**

In cooperation with
**AAAI – Association for the Advancement of Artificial Intelligence**

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

Printed in Portugal

ISBN: 978-989-8111-32-6

Depósito Legal: 273830/08

http://www.icinco.org

secretariat@icinco.org

# BRIEF CONTENTS

# INVITED SPEAKERS

**Miguel Ayala Botto**

Instituto Superior Técnico

Portugal

**Peter Simon Sapaty**

Institute of Mathematical Machines and Systems

National Academy of Sciences

Ukraine

**Ronald C. Arkin**

Georgia Institute of Technology

U.S.A.

**Marco Dorigo**

IRIDIA, Université Libre de Bruxelles

Belgium

# ORGANIZING AND STEERING COMMITTEES

### CONFERENCE CO-CHAIRS

Jorge Cardoso, University of Madeira (UMa), Madeira, Portugal

Joaquim Filipe, INSTICC / Polytechnic Institute of Setúbal, Portugal

### PROGRAM CO-CHAIRS

Juan Andrade Cetto, Universitat Autònoma de Barcelona, Spain

Jean-Louis Ferrier, University of Angers, France

### LOCAL ARRANGEMENTS

Laura Rodriguez, University of Madeira (UMa), Portugal

### PROCEEDINGS PRODUCTION

Andreia Costa, INSTICC, Portugal

Bárbara Morais, INSTICC, Portugal

Bruno Encarnação, INSTICC, Portugal

Helder Coelhas, INSTICC, Portugal

Paulo Brito, INSTICC, Portugal

Vera Coelho, INSTICC, Portugal

Vera Rosário, INSTICC, Portugal

Vitor Pedrosa, INSTICC, Portugal

### CD-ROM PRODUCTION

Elton Mendes, INSTICC, Portugal

### WEBDESIGNER AND GRAPHICS PRODUCTION

Marina Carvalho, INSTICC, Portugal

### SECRETARIAT AND WEBMASTER

Marina Carvalho, INSTICC, Portugal

# PROGRAM COMMITTEE

**Arturo Hernandez Aguirre**, Centre for Research in Mathematics, Mexico

**Eugenio Aguirre**, University of Granada, Spain

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

**Frank Allgower**, University of Stuttgart, Germany

**Fouad Al-Sunni**, KFUPM, Saudi Arabia

**Bala Amavasai**, Sheffield Hallam University, U.K.

**Francesco Amigoni**, Politecnico di Milano, Italy

**Yacine Amirat**, University Paris 12, France

**Nicolas Andreff**, LASMEA, France

**Stefan Andrei**, Lamar University, U.S.A.

**Plamen Angelov**, Lancaster University, U.K.

**Luis Antunes**, GUESS/Universidade de Lisboa, Portugal

**Peter Arato**, Budapest University of Technology and Economics, Hungary

**Helder Araújo**, University of Coimbra, Portugal

**Gustavo Arroyo-Figueroa**, Instituto de Investigaciones Electricas, Mexico

**Marco Antonio Arteaga**, Universidad Nacional Autonoma de Mexico, Mexico

**Vijanth Sagayan Asirvadam**, University Technology Petronas, Malaysia

**Wudhichai Assawinchaichote**, King Mongkut's University of Technology Thonburi, Thailand

**Robert Babuska**, TU Delft, The Netherlands

**Ruth Bars**, Budapest University of Technology and Economics, Hungary

**Adil Baykasoglu**, University of Gaziantep, Turkey

**Laxmidhar Behera**, Indian Institute of Technology, India

**Maren Bennewitz**, University of Freiburg, Germany

**Karsten Berns**, University Kaiserslautern, Germany

**Arijit Bhattacharya**, The Patent Office, India

**Robert Bicker**, Newcastle University, U.K.

**Sergio Bittanti**, Politecnico Di Milano, Italy

**Stjepan Bogdan**, University of Zagreb, Croatia

**Jean-Louis Boimond**, LISA, France

**Djamel Bouchaffra**, Grambling State University, U.S.A.

**Patrick Boucher**, SUPELEC, France

**Guy Boy**, European Institute of Cognitive Sciences and Engineering (EURISCO International), France

**Bernard Brogliato**, INRIA, France

**Edmund Burke**, University of Nottingham, U.K.

**Kevin Burn**, University of Sunderland, U.K.

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

**Dídac Busquets**, Universitat de Girona, Spain

**Luis M. Camarinha-Matos**, New University of Lisbon, Portugal

**Marc Carreras**, University of Girona, Spain

**Jorge Martins de Carvalho**, FEUP, 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

**YangQuan Chen**, Utah State University, U.S.A.

**Albert M. K. Cheng**, University of Houston, U.S.A.

**Graziano Chesi**, University of Hong Kong, China

**Yiu-ming Cheung**, Hong Kong Baptist University, Hong Kong

**Sung-Bae Cho**, Yonsei University, Korea

**Ryszard S. Choras**, University of Technology & Agriculture, Poland

**Carlos Coello Coello**, CINVESTAV-IPN, Mexico

**Patrizio Colaneri**, Politecnico di Milano, Italy

**António Dourado Correia**, University of Coimbra, Portugal

**Yechiel Crispin**, Embry-Riddle University, U.S.A.

**Danilo De Rossi**, University of Pisa, Italy

**Elena De Santis**, University of L'Aquila, Italy

**Matthias Dehmer**, TU Vienna, Austria

**Angel P. del Pobil**, Universitat Jaume I, Spain

**Mingcong Deng**, Okayama University, Japan

# PROGRAM COMMITTEE (CONT.)

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

**Jorge Dias**, ISR - Institute of Systems and Robotics, Portugal

**Rüdiger Dillmann**, University of Karlsruhe, Germany

**Denis Dochain**, Université Catholique de Louvain, Belgium

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

**Alexandre Dolgui**, Ecole des Mines de Saint Etienne, France

**Marco Dorigo**, Université Libre de Bruxelles, Belgium

**Petr Ekel**, Pontifical Catholic University of Minas Gerais, Brazil

**Sebastian Engell**, TU Dortmund, Germany

**Simon Fabri**, University of Malta, Malta

**Sergej Fatikow**, University of Oldenburg, Germany

**Jean-Marc Faure**, Ecole Normale Superieure de Cachan, France

**Jean-Louis Ferrier**, Université d'Angers, France

**Limor Fix**, Intel, U.S.A.

**Juan F. Flores**, University of Michoacan, Mexico

**Georg Frey**, University of Kaiserslautern, Germany

**Manel Frigola**, Thecnical University of Catalonia (UPC), Spain

**Colin Fyfe**, University of Paisley, U.K.

**Dragan Gamberger**, Rudjer Boskovic Institute, Croatia

**Leonardo Garrido**, Tecnológico de Monterrey, Mexico

**Nicholas Gans**, University of Florida, U.S.A.

**Ryszard Gessing**, Silesian University of Technology, Poland

**Lazea Gheorghe**, Technical University of Cluj-Napoca, Romania

**Maria Gini**, University of Minnesota, U.S.A.

**Alessandro Giua**, University of Cagliari, Italy

**Luis Gomes**, Universidade Nova de Lisboa, Portugal

**John Gray**, University of Salford, U.K.

**Dongbing Gu**, University of Essex, U.K.

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

**Jason Gu**, Dalhousie University, Canada

**José J. Guerrero**, Universidad de Zaragoza, Spain

**Jatinder (Jeet) Gupta**, University of Alabama in Huntsville, U.S.A.

**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 Electronics & Safety, U.S.A.

**Uwe D. Hanebeck**, Universität Karlsruhe (TH), Germany

**John Harris**, University of Florida, U.S.A.

**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

**Alamgir Hossain**, Bradford University, U.K.

**Dimitrios Hristu-Varsakelis**, University of Macedonia, Greece

**Guoqiang Hu**, University of Florida, U.S.A.

**Nor Ashidi Mat Isa**, Universiti Sains Malaysia, Malaysia

**Ray Jarvis**, Monash University, Australia

**Odest Jenkins**, Brown University, U.S.A.

**Ping Jiang**, The University of Bradford, U.K.

**Agustin Jimenez**, Universidad Politécnica de Madrid, Spain

**Ivan Kalaykov**, Örebro University, Sweden

**Michail Kalogiannakis**, University Paris 5 - René Descartes, France

**Dimitrios Karras**, Chalkis Institute of Technology, Greece

**Fakhri Karray**, University of Waterloo, Canada

**Dusko Katic**, Mihailo Pupin Institute, Serbia

**Graham Kendall**, The University of Nottingham, U.K.

# PROGRAM COMMITTEE (CONT.)

**Bart Kosko**, University of Southern California, U.S.A.

**George L. Kovács**, Hungarian Academy of Sciences, Hungary

**Krzysztof Kozlowski**, Poznan University of Technology, Poland

**Gerhard Kraetzschmar**, Bonn-Rhein-Sieg University of Applied Sciences, Germany

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

**Cecilia Laschi**, Scuola Superiore Sant'Anna, Italy

**Jean-Claude Latombe**, Stanford University, U.S.A.

**M. Kemal Leblebicioglu**, Middle East Technical University, Turkey

**Loo Hay Lee**, National University of Singapore, Singapore

**Soo-Young Lee**, KAIST, Korea

**Graham Leedham**, University of New South Wales, Singapore

**Kauko Leiviskä**, University of Oulu, Finland

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

**Yangmin Li**, University of Macau, China

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

**Vincenzo Lippiello**, Università Federico II di Napoli, Italy

**Honghai Liu**, University of Portsmouth, U.K.

**Luís Seabra Lopes**, Universidade de Aveiro, Portugal

**Brian Lovell**, The University of Queensland, Australia

**Peter Luh**, University of Connecticut, U.S.A.

**Jose Tenreiro Machado**, Institute of Engineering of Porto, Portugal

**Anthony Maciejewski**, Colorado State University, U.S.A.

**N. P. Mahalik**, California State University, Fresno, U.S.A.

**Bruno Maione**, Politecnico di Bari, Italy

**Frederic Maire**, Queensland University of Technology, Australia

**Om Malik**, University of Calgary, Canada

**Jacek Mandziuk**, Warsaw University of Technology, Poland

**Hervé Marchand**, INRIA, France

**Philippe Martinet**, LASMEA, France

**Aníbal Matos**, Faculdade de Engenharia da Universidade do Porto (FEUP), Portugal

**Rene V. Mayorga**, University of Regina, Canada

**Barry McCollum**, Queen's University Belfast, U.K.

**Ken McGarry**, University of Sunderland, U.K.

**Gerard McKee**, The University of Reading, U.K.

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

**Patrick Millot**, Université de Valenciennes, France

**José Mireles Jr.**, Universidad Autonoma de Ciudad Juarez, Mexico

**Masoud Mohammadian**, University of Canberra, Australia

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

**Vladimir Mostyn**, VSB - Technical University of Ostrava, Czech Republic

**Rafael Muñoz-Salinas**, University of Cordoba, Spain

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

**Fazel Naghdy**, University of Wollongong, Australia

**Tomoharu Nakashima**, Osaka Prefecture University, Japan

**Andreas Nearchou**, University of Patras, Greece

**Luciana Porcher Nedel**, Universidade Federal do Rio Grande do Sul (UFRGS), Brazil

**Sergiu Nedevschi**, Technical University of Cluj-Napoca, Romania

**Maria Neves**, Instituto Superior de Engenharia do Porto, Portugal

**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

**Tsukasa Ogasawara**, Nara Institute of Science and Technology, Japan

# PROGRAM COMMITTEE (CONT.)

**José Valente de Oliveira**, Universidade do Algarve, Portugal

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

**Djamila Ouelhadj**, University of Nottingham, ASAP GROUP (Automated Scheduling, Optimisation and Planning), U.K.

**Christos Panayiotou**, University of Cyprus, Cyprus

**Stefano Panzieri**, Universitá degli Studi "Roma Tre", Italy

**Evangelos Papadopoulos**, NTUA, Greece

**Michel Parent**, INRIA, France

**Igor Paromtchik**, RIKEN, Japan

**Mario Pavone**, University of Catania, Italy

**Witold Pedrycz**, University of Alberta, Canada

**Carlos Eduardo Pereira**, Federal University of Rio Grande do Sul - UFRGS, Brazil

**Duc Pham**, Cardiff University, U.K.

**J. Norberto Pires**, University of Coimbra, Portugal

**Marios Polycarpou**, University of Cyprus, Cyprus

**Marie-Noëlle Pons**, CNRS, France

**Raul Marin Prades**, Jaume I University, Spain

**Libor Preucil**, Czech Technical University in Prague, Czech Republic

**José Ragot**, Institut National Polytechnique de Lorraine, France

**A. Fernando Ribeiro**, Universidade do Minho, Portugal

**Robert Richardson**, University of Manchester, U.K.

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

**Kurt Rohloff**, BBN Technologies, U.S.A.

**Juha Röning**, University of Oulu, Finland

**Agostinho Rosa**, IST, Portugal

**António Ruano**, CSI, Portugal

**Fariba Sadri**, Imperial College London, U.K.

**Carlos Sagüés**, University of Zaragoza, Spain

**Mehmet Sahinkaya**, University of Bath, U.K.

**Priti Srinivas Sajja**, Sardar Patel University, India

**Antonio Sala**, Universidad Politecnica de Valencia, Spain

**Abdel-Badeeh Salem**, Ain Shams University, Egypt

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

**Nilanjan Sarkar**, Vanderbilt University, U.S.A.

**Jurek Sasiadek**, Carleton University, Canada

**Daniel Sbarbaro**, Universidad de Concepcion, Chile

**Carsten Scherer**, Delft University of Technology, The Netherlands

**Matthias Scheutz**, Indiana University, U.S.A.

**Klaus Schilling**, University Würzburg, Germany

**Carla Seatzu**, University of Cagliari, Italy

**Rodolphe Sepulchre**, University of Liege, Belgium

**Michael Short**, University of Leicester, U.K.

**Bruno Siciliano**, Università di Napoli Federico II, Italy

**João Silva Sequeira**, Instituto Superior Técnico, Portugal

**Silvio Simani**, University of Ferrara, Italy

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

**Michael Small**, Hong Kong Polytechnic University, China

**Cyrill Stachniss**, University of Freiburg, Germany

**Burkhard Stadlmann**, University of Applied Sciences Wels, Austria

**Tarasiewicz Stanislaw**, Université Laval, Canada

**Olaf Stursberg**, Technical University of Munich, Germany

**Chun-Yi Su**, Concordia University, Canada

**Raúl Suárez**, Universitat Politecnica de Catalunya (UPC), Spain

**Ryszard Tadeusiewicz**, AGH University of Science and Technology, Poland

**Tianhao Tang**, Shanghai Maritime University, China

**Adriana Tapus**, University of Southern California, U.S.A.

**József K. Tar**, Budapest Tech Polytechnical Institution, Hungary

**Daniel Thalmann**, EPFL, Switzerland

# PROGRAM COMMITTEE (CONT.)

**Gui Yun Tian**, University of Newcastle, U.K.

**Avgoustos Tsinakos**, T.E.I. Kavalas, Greece

**Antonios Tsourdos**, Cranfield University, U.K.

**Nikos Tsourveloudis**, Technical University of Crete, Greece

**Ivan Tyukin**, University of Leicester, U.K.

**Anthony Tzes**, University of Patras, Greece

**Masaru Uchiyama**, Tohoku University, Japan

**Dariusz Ucinski**, University of Zielona Gora, Poland

**Nicolas Kemper Valverde**, Universidad Nacional Autónoma de México, Mexico

**Marc Van Hulle**, K. U. Leuven, Belgium

**Gerrit van Straten**, Wageningen University, Netherlands

**Eloisa Vargiu**, University of Cagliari, Italy

**Annamaria R. Varkonyi-Koczy**, Budapest University of Technology and Economics, Hungary

**Laurent Vercouter**, Ecole des Mines de Saint-Etienne, France

**Luigi Villani**, Università di Napoli Federico II, Italy

**Bernardo Wagner**, University of Hannover, Germany

**Axel Walthelm**, sepp.med GmbH, Germany

**Dianhui Wang**, La Trobe University, Australia

**Lipo Wang**, Nanyang Technological University, Singapore

**Zidong Wang**, Brunel University, U.K.

**Vincent Wertz**, Université catholique de Louvain, Belgium

**Dirk Wollherr**, Technische Universität München, Germany

**Sangchul Won**, Pohang University of Science and Technology, Korea

**Peter Xu**, Massey University, New Zealand

**Bin Yao**, Purdue University, U.S.A.

**Xinghuo Yu**, Royal Melbourne Institute of Techology, Australia

**Marek Zaremba**, Universite du Quebec, Canada

**Janan Zaytoon**, University of Reims Champagne Ardenne, France

**Du Zhang**, California State University, U.S.A.

**Changjiu Zhou**, Singapore Polytechnic, Singapore

**Dayong Zhou**, Cirrus Logic Inc., U.S.A.

**Primo Zingaretti**, Università Politecnica delle Marche, Italy

**Argyrios Zolotas**, Loughborough University, U.K.

# AUXILIARY REVIEWERS

**Hyo-Sung Ahn**, Gwangju Institute of Science and Technology, Korea

**Prasanna Balaprakash**, IRIDIA, CoDE, Université Libre de Bruxelles, Belgium

**Majid Chauhdry**, University of Connecticut, U.S.A.

**Ying Chen**, University of Connecticut, U.S.A.

**Pedro Fernandes**, Institute of Systems and Robotics, UC, Portugal

**Matteo De Felice**, Univ. Roma TRE, Italy

**Michele Folgheraiter**, German Research Center for Artificial Intelligence, Germany

**Jun Fu**, Concordia University, Canada

**Andrea Gasparri**, University Roma TRE, Italy

**Emmanuel Godoy**, Supelec, France

**Che Guan**, University of Connecticut, U.S.A.

**Istvan Harmati**, Budapest University of Technology and Economics, Hungary

**Abhinaya Joshi**, University of Connecticut, U.S.A.

**Balint Kiss**, Budapest University of Technology and Economics, Hungary

**Gabor Kovacs**, Budapest University of Technology and Economics, Hungary

**Roland Lenain**, Cemagref, France

**Nikolay Manyakov**, K. U. Leuven, Germany

**Philippe Martinet**, LASMEA, Blaise Pascal University, France

**Sandro Meloni**, University Roma TRE, Italy

**Eduardo Montijano Muñoz**, University of Zaragoza, Spain

**A. C. Murillo**, Universidad de Zaragoza, Spain

**Gonzalo Lopez Nicolas**, University of Zaragoza, Spain

**Sorin Olaru**, Supelec, France

**Federica Pascucci**, Univ. Roma TRE, Italy

**Karl Pauwels**, K. U. Leuven, Germany

**Paulo Peixoto**, University of Coimbra, Portugal

**Jun Peng**, Chongqing University of Science and Technology, China

**Luis Puig**, Universidad de Zaragoza, Spain

**Maurizio di Rocco**, University Roma, TRE, Italy

**Marco Montes De Oca Roldan**, IRIDIA, CoDE, Université Libre de Bruxelles, Belgium

**Joerg Stueckler**, University of Freiburg, Germany

**Jin Sun**, Tsinghua University, China

**Emese Szadeczky-Kardos**, Budapest University of Technology and Economics, Hungary

**Sihem Tebbani**, Supelec, France

**Benoit Thuilot**, LASMEA, Blaise Pascal University, France

**Guoyu Tu**, Tsinghua University, Beijing, China

**Peng Wang**, University of Connecticut, U.S.A.

**Weihua Wang**, University of Connecticut, U.S.A.

**Bingjie Zhang**, University of Connecticut, U.S.A.

**Yige Zhao**, University of Connecticut, U.S.A.

**Ying Zhao**, University of Connecticut, U.S.A.

# SELECTED PAPERS BOOK

A number of selected papers presented at ICINCO 2008 will be published by Springer-Verlag in a LNEE Series book. This selection will be done by the Conference Co-chairs and Program Co-chairs, among the papers actually presented at the conference, based on a rigorous review by the ICINCO 2008 program committee members.

# FOREWORD

This book contains the proceedings of the 5th International Conference on Informatics in Control, Automation and Robotics (ICINCO 2008) which was organized by the Institute for Systems and Technologies of Information, Control and Communication (INSTICC) in collaboration with the University of Madeira (UMa) and held in Madeira. ICINCO 2008 was technically co-sponsored by the IEEE Systems Man and Cybernetics Society (IEEE-SMC) and 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 four plenary keynote lectures, given by internationally recognized researchers, namely - Miguel A. Botto (Instituto Superior Técnico, Portugal), Peter S. Sapaty (Institute of Mathematical Machines and Systems, National Academy of Sciences, Ukraine), Ronald C. Arkin (Georgia Institute of Technology, U.S.A.), and Marco Dorigo (IRIDIA, Université Libre de Bruxelles, Belgium). These keynote speakers participated also on a plenary panel entitled "*The new frontiers of Control, Automation and Robotics*".

The meeting is complemented with two satellite workshops and two special sessions, 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 Control Systems (IVCS), the Special Session on Service Oriented Architectures for SME robots and Plug-and-Produce, and the Special Session on Multi-Agent Robotic Systems.

ICINCO received 392 paper submissions, not including those of workshops and special sessions, from more than 50 countries, in all continents. To evaluate each submission, a double blind paper review was performed by the Program Committee, whose members are highly qualified researchers in ICINCO topic areas. Finally, only 190 papers are published in these proceedings and presented

at the conference. Of these, 114 papers were selected for oral presentation (33 full papers and 81 short papers) and 76 papers were selected for poster presentation. The full paper acceptance ratio was 8,4%, and the oral acceptance ratio (including full papers and short papers) was 29%. 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 2008.

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 Welcome Drink to all participants in the afternoon of May 11 (Sunday) and a Conference and Workshops Social Event & Banquet in the evening of May 14 (Wednesday).

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 the 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 2009, to be held in July in Milan.

**Joaquim Filipe**
Polytechnic Institute of Setúbal / INSTICC, Portugal

**Juan Andrade-Cetto**
Institut de Robotica i Informatica Industrial, CSIC-UPC, Spain

**Jean-Louis Ferrier**
LISA-ISTIA – Université d'Angers, France

# CONTENTS

**SHORT PAPERS**

**POSTERS**

# INVITED SPEAKERS

# KEYNOTE LECTURES

# DEALING WITH UNCERTAINTY IN THE HYBRID WORLD*

Luís Pina and Miguel Ayala Botto
*Department of Mechanical Engineering, IDMEC, Instituto Superior Técnico*
*Technical University of Lisbon, Portugal*
*luispina@dem.ist.utl.pt, ayalabotto@ist.utl.pt*

Keywords: Hybrid systems, Hybrid Estimation, Interacting multiple-model estimation, Observability.

Abstract: This paper presents an efficient state estimation algorithm for hybrid systems based on a least-squares Interacting Multiple-Model setup. The proposed algorithm is shown to be computationally efficient when compared with the Moving Horizon Estimaton algorithm that is a brute force optimization algorithm for simultaneous discrete mode and continuous state estimation of a hybrid system. The main reason has to do with the fact that the proposed algorithm is able to disregard as many discrete mode sequence estimates as possible. This is done by rapidly computing good estimates, separating the constrained and unconstrained estimates, and using some auxiliary coefficients computed off-line. The success of this state estimation algorithm is shown for a fault detection problem of the benchmark AMIRA DTS200 three-tanks system experimental setup.

## 1 INTRODUCTION

In the last decade hybrid systems have become a major research topic in Control Engineering (Antsaklis, 2000). Hybrid systems are dynamical systems composed by both discrete valued and continuous valued states. The dynamics of a hybrid system is governed by a mode selector that determines, at each time instant, which discrete mode is active from endogenous and/or exogenous variables. The continuous state is then updated through a dynamic relation that is selected from a set of possible dynamics according to the value of the active discrete mode. In fact, the presence of physical components such as on/off switches or valves, gears or speed selectors, or behaviors dependent on if-then-else rules imply explicitly or implicitly the discrete/continuous interaction. This interaction can be found in many real world applications such as automotive control, urban and air traffic control, communications networks, embedded control systems, and in the control of complex industrial systems via the combination of classical continuous control laws with supervisory switching logic.

The hybrid nature has attracted the interest of mathematicians, control engineers and computer scientists, therefore leading to different modeling languages and paradigms that influenced the line of research on hybrid systems in several different ways. For instance, the computer science research community is more focused on systems whose variables take values in a finite set, so adopted the discrete events modeling formalism to model hybrid systems, using finite state machines, Petri nets, temporal logic, etc. On the other hand, the control systems community typically considers a continuous valued world, where time is continuously changing, thus considering a hybrid system as described by a differential (or difference) equation with some switching mechanism. Examples of such hybrid models include Piece-Wise Affine (PWA) (Sontag, 1981) and Mixed Logical Dynamical (MLD) (Bemporad and Morari, 1999) models. A PWA model is the most intuitive representation of a hybrid system since it provides a direct relation to linear systems while still capturing very complex dynamical behaviors. However, a MLD representation is most adequate to be used in optimization problems since it is able to embed both propositional logic statements (if-then-else rules) and operating constraints in a state linear dynamics equation by transforming them to mixed-integer linear inequalities. Despite these differences, PWA and MLD are equivalent models of hybrid systems in respect to well-posedness and boundness of input, state, output or auxiliary variables (Heemels et al., 2001). This fact allows to interchange analysis and synthesis tools between them.

*This work was supported by project PTDC/EME-CRO/69117/2006 co-sponsored by FEDER, Programa Operacional Ciência e Inovação 2010, Portugal.

Research on hybrid systems spans to a wide range of topics (and approaches), from modeling to stability analysis, reachability analysis and verification, study of the observability and controllability properties, methods of state estimation and fault detection, identification techniques, and control methodologies. Typically, hybrid tools rely on the solution of optimization problems. However, due to the different nature of the optimization variables involved (integer and continuous) the main source of complexity becomes the combinatorial (yet finite) number of possible switching sequences that have to be considered. A hybrid optimal solution thus requires solving mixed-integer non-convex optimization algorithms with NP-complete complexity (Torrisi and Bemporad, 2001).

Analysis and synthesis procedures for hybrid systems when disturbances are present either on the continuous dynamics or on the discrete mode of the hybrid system, is still an open research topic that has being tackled by several authors using distinct approaches. In the state estimation problem two distinct approaches are usually followed, the main difference being the knowledge of the active mode: some approaches consider only continuous state uncertainty with known discrete mode, while others assume that both the discrete mode and the continuous state are unknown. The combination of both uncertainties (state and mode) on the estimation process of a hybrid system presents a very difficult problem for which a global solution is not yet found. When the discrete mode is known in advance, the problem is greatly simplified and the state estimation methodologies for linear systems can be applied with very little modifications. For example in (Böker and Lunze, 2002) a bank of Kalman filters is used and in (Alessandri and Coletta, 2003) an LMI based algorithm computes the stabilizing gains for a set of Luenberger observers. If, on the other hand, the discrete mode must also be estimated the estimation problem becomes much more complex and every discrete mode sequence (*dms*) must be checked to choose the one that provides the best fit for the observed data. The continuous state estimates are then computed for the estimated *dms*. Several works address this problem, see (Balluchi et al., 2002) where a location observer is used to estimate the discrete mode and a Luenberger observer is then used to estimate the continuous state. In (Ferrari-Trecate et al., 2002) and (Pina and Botto, 2006) a Moving Horizon Estimation (MHE) scheme simultaneously estimates the discrete mode and the continuous state, differing in the fact that the latter can also estimate the input disturbances.

The derivation of the truly optimal filter for systems with switching parameters was first presented in (Athans and Chang, 1976). The objective was to perform simultaneous system identification and state estimation for linear systems but the derivation is quite general and is directly applicable to the hybrid state estimation problem. This method requires the consideration of all admissible *dms* starting from the initial time instant, being obviously unpractical since the number of *dms* grows exponentially in time, and so, suboptimal methods were developed. From the various possibilities, considering all the admissible *dms* of a given length is usually the preferred methodology. In view of this, suboptimal multiple model estimation schemes where then developed and applied for tracking maneuvering vehicles, as surveyed in (Mazor et al., 1998), and systems with Markovian switching coefficients, (Blom and Bar-Shalom, 1988), proving their efficiency for state estimation in multiple model systems. Multiple model estimation algorithms use a set of filters, one for each possible dynamic of the system. In this paper an efficient state estimation algorithm for stochastic hybrid systems, based on the Interacting Multiple-Model (IMM) estimation algorithm, is proposed. The method is applicable to most of the existing models of hybrid systems subject to disturbances with explicitly known probability density function, so being rather general. This estimation method will be further compared to the Moving Horizon Estimation (MHE) algorithm and tested in the benchmark AMIRA DTS200 three-tanks system experimental setup.

The paper is organized as follows. Section 2 provides a description of the considered PWA model and in section 3 the proposed Interacting Multiple-Model estimation algorithm is presented. Section 4 presents an experimental application of the proposed algorithms to the AMIRA DTS200 three-tanks system experimental setup. First the experimental setup is presented and modelled, including a full characterization of all uncertainties. Then the proposed algorithms are tested and their performance is compared. Finally, in section 5 some conclusions are drawn along with some possible future developments.

## 2 SYSTEM DESCRIPTION

The proposed estimation algorithm is developed for PWA systems which were introduced in (Sontag, 1981). The following stochastic PWA model will be considered:

$$x(k+1) = A_{i(k)}x(k)+B_{i(k)}u(k)+f_{i(k)}+L_{i(k)}w(k) \tag{1a}$$

$$y(k) = C_{i(k)}x(k)+D_{i(k)}u(k)+g_{i(k)}+v(k) \tag{1b}$$

$$\text{iff } \begin{bmatrix} x(k) \\ u(k) \\ w(k) \end{bmatrix} \in \Omega_{i(k)} \tag{1c}$$

where $k$ is the discrete time, $x(k) \in \mathbb{X} \subset \mathbb{R}^{n_x}$ is the continuous state, $u(k) \in \mathbb{U} \subset \mathbb{R}^{n_u}$ is the input, $y(k) \in \mathbb{R}^{n_y}$ is the output, $i(k) \in I = \{1,\ldots,s\}$ is the discrete mode, and $s$ is the total number of discrete modes. The matrices and vectors $A_i$, $B_i$, $f_i$, $L_i$, $C_i$, $D_i$, $g_i$ depend on the discrete mode $i(k)$ and have appropriate dimensions. The input disturbance $w(k)$ and the measurement noise $v(k)$ are modelled as independent identically distributed random variables, belonging to the sets $\mathbb{W}_i$ and $\mathbb{V}_i$, with expected values $E\{w(k)\} = 0$, $E\{v(k)\} = 0$ and covariances $\Sigma_{w_i}$ and $\Sigma_{v_i}$, respectively. These conditions are not restrictive at all since the zero mean can be imposed by summing a constant vector to the disturbances and compensated in the affine term of the system dynamics (1) and, the sets $\mathbb{W}_i$ and $\mathbb{V}_i$ can be considered large enough to contain all possible disturbances relevant for practical applications, for instance 99.99% of all admissible values. Notice that the input disturbance and measurement noise *pdf*s may depend on the actual mode of the system $i(k)$. The sets $\mathbb{W}_i$ and $\mathbb{V}_i$ are respectively defined for each mode $i(k)$ by:

$$H_{\mathbb{W}_{i(k)}}\, w(k) \leq h_{\mathbb{W}_{i(k)}} \quad , \quad \forall_{k\in\mathbb{N}_0} \tag{2}$$

$$H_{\mathbb{V}_{i(k)}}\, v(k) \leq h_{\mathbb{V}_{i(k)}} \quad , \quad \forall_{k\in\mathbb{N}_0} \tag{3}$$

The discrete mode $i(k)$ is a piecewise constant function of the state, input and input disturbance of the system whose value is defined by the regions $\Omega_i$:

$$\Omega_i \quad : \quad S_i\, x(k) + R_i\, u(k) + Q_i\, w(k) \leq T_i \tag{4}$$

Some helpful notation regarding the time-compressed representation of (Kamen, 1992) for system (1) will now be introduced. The time-compressed representation of a system defines the dynamics of the system over a sequence of time instants in opposition to the single time step state-space representation. Consider the time interval $[k, k+T-1]$, the sequence of discrete modes over this interval is represented as $\mathbf{i}_T = \mathbf{i}_T(k) \triangleq \{i(k),\ldots,i(k+T-1)\}$. To simplify the notation, the time index $k$ is removed from the discrete mode sequence (*dms*) whenever it is obvious from the other elements in the equations. In view of this, the output sequence over the same interval can be computed by:

$$Y_T(k) = \mathbf{C}_{\mathbf{i}_T}x(k)+\mathbf{D}_{\mathbf{i}_T}U_T(k)+\mathbf{g}_{\mathbf{i}_T}+\mathbf{L}_{\mathbf{i}_T}W_T(k)+V_T(k) \tag{5}$$

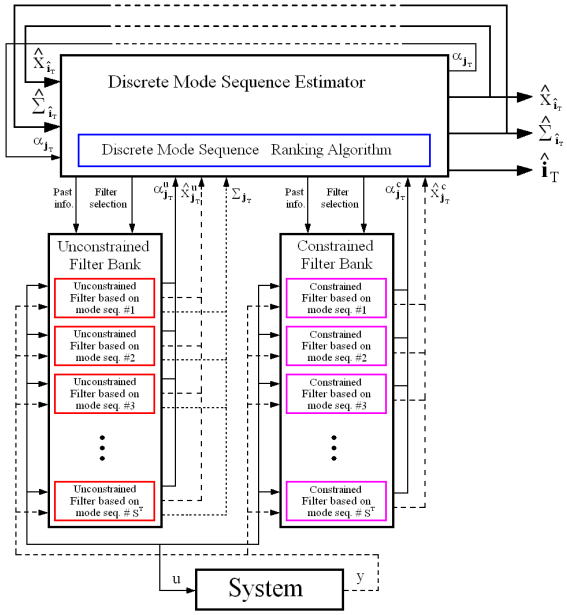



Figure 1: Interacting Multiple-Model Estimation Algorithm.

where the input, input disturbance and measurement noise sequences $U_T(k)$, $W_T(k)$ and $V_T(k)$ respectively are defined in the same way as the output sequence $Y_T(k) \triangleq [y(k)^{\mathrm{T}},\ldots,y(k+T-1)^{\mathrm{T}}]^{\mathrm{T}}$. The matrices and vectors $\mathbf{C}_{\mathbf{i}_T}$, $\mathbf{D}_{\mathbf{i}_T}$, $\mathbf{g}_{\mathbf{i}_T}$ and $\mathbf{L}_{\mathbf{i}_T}$ are computed from the system dynamics (1a-1b) according to what is presented in (Kamen, 1992). The same reasoning can be applied to the constraints $\Omega_{\mathbf{i}_T}$:

$$\Omega_{\mathbf{i}_T} : \quad \mathbf{S}_{\mathbf{i}_T}\, x(k) + \mathbf{R}_{\mathbf{i}_T}\, U_T(k) + \mathbf{Q}_{\mathbf{i}_T}\, W_T(k) \leq \mathbf{T}_{\mathbf{i}_T} \tag{6}$$

where the matrices $\mathbf{S}_{\mathbf{i}_T}$, $\mathbf{R}_{\mathbf{i}_T}$, $\mathbf{Q}_{\mathbf{i}_T}$ and $\mathbf{T}_{\mathbf{i}_T}$ can be computed from the system dynamics (1a) and partitions (4). The inequalities that define the disturbance and noise sets over a *dms* $\mathbf{i}_T$, $\mathbb{W}_{\mathbf{i}_T}$ and $\mathbb{V}_{\mathbf{i}_T}$ respectively, can also be easily found from equations (2) and (3):

$$\mathbf{H}_{\mathbb{W}_{\mathbf{i}_T}}\, W_T(k) \leq \mathbf{h}_{\mathbb{W}_{\mathbf{i}_T}} \tag{7}$$

$$\mathbf{H}_{\mathbb{V}_{\mathbf{i}_T}}\, V_T(k) \leq \mathbf{h}_{\mathbb{V}_{\mathbf{i}_T}} \tag{8}$$

# 3 INTERACTING MULTIPLE MODEL ESTIMATION

The proposed Interacting Multiple-Model (IMM) Estimation algorithm is composed of three parts; the Unconstrained Filter Bank (UFB), the Constrained Filter Bank (CFB) and, the Discrete Mode Sequence Estimator (DMSE). A schematic representation is presented in figure 1.

The estimation algorithm works as follows: first the continuous state estimates are computed in the UFB without considering the constraints. Then, the DMSE computes the squared errors of these estimates and ranks them. Finally, starting with the estimate with the lowest squared error, the estimates are recomputed in the CFB considering the presence of constraints. When the most accurate estimate is already a constrained estimate the whole process stops.

As the estimation is based on sequences of measurements $Y_T(k)$ and discrete modes $\mathbf{i}_T(k)$, two distinct time instants must be considered: the time instant at the beginning of the sequences, $k$, and the time instant at the end of these sequences, which is the present time instant $t = k+T-1$. The state estimates will be computed at time instant $k$, and can be propagated to the present time instant according to the estimated dynamics.

## 3.1 Unconstrained Filter Bank

The UFB computes the unconstrained state estimates. It is composed by a set of unconstrained least-squares filters, one for each possible *dms* $\mathbf{j}_T$:

$$\hat{x}^u_{\mathbf{j}_T}(k|t) = \hat{x}_{\mathbf{j}_T}(k|t-1) + \tag{9}$$

$$\mathbf{K}_{\mathbf{j}_T}(k|t-1)\left[\left(Y_T(k)-\mathbf{D}_{\mathbf{j}_T}U_T(k)-\mathbf{g}_{\mathbf{j}_T}\right)-\mathbf{C}_{\mathbf{j}_T}\hat{x}_{\mathbf{j}_T}(k|t-1)\right]$$

where $\hat{x}_{\mathbf{j}_T}(k|t-1)$ is the *a priori* continuous state estimate for mode sequence $\mathbf{j}_T$ using measurements up to time instant $t-1$. $\mathbf{K}_{\mathbf{j}_T}(k|t-1)$ is the filter gain:

$$\mathbf{K}_{\mathbf{j}_T}(k|t-1) = \left(\Sigma^{-1}_{x_{\mathbf{j}_T}}(k|t-1)+\mathbf{C}^{\mathrm{T}}_{\mathbf{j}_T}\Sigma^{-1}_{Y_{\mathbf{j}_T}}\mathbf{C}_{\mathbf{j}_T}\right)^{-1}\mathbf{C}^{\mathrm{T}}_{\mathbf{j}_T}\Sigma^{-1}_{Y_{\mathbf{j}_T}} \tag{10}$$

$$\Sigma_{Y_{\mathbf{j}_T}} = \left[\mathbf{L}_{\mathbf{j}_T}\ I_{T.n_y}\right]\begin{bmatrix}\Sigma_{W_{\mathbf{j}_T}} & 0\\ 0 & \Sigma_{V_{\mathbf{j}_T}}\end{bmatrix}\left[\mathbf{L}_{\mathbf{j}_T}\ I_{T.n_y}\right]^{\mathrm{T}} \tag{11}$$

The covariance of the obtained unconstrained estimate can also be computed:

$$\Sigma_{x_{\mathbf{j}_T}}(k|t) = \left(\Sigma^{-1}_{x_{\mathbf{j}_T}}(k|t-1)+\mathbf{C}^{\mathrm{T}}_{\mathbf{j}_T}\Sigma^{-1}_{Y_{\mathbf{j}_T}}\mathbf{C}_{\mathbf{j}_T}\right)^{-1} \tag{12}$$

This covariance matrix not only provides some insight on the accuracy of the continuous state estimate $\hat{x}^u_{\mathbf{j}_T}(k|t)$, but also defines the confidence on the past information at the subsequent time instant $\hat{x}_{\mathbf{j}_T}(k+1|t)$:

$$\Sigma_{x_{\mathbf{j}_T}}(k+1|t) = A_{j(k)}\Sigma_{x_{\mathbf{j}_T}}(k|t)A^{\mathrm{T}}_{j(k)} + L_{j(k)}\Sigma_{wj(k)}L^{\mathrm{T}}_{j(k)} \tag{13}$$

When computing the unconstrained state estimate, no *a priori* information may be available or one may be interested in discarding it, then $\Sigma^{-1}_{x_{\mathbf{j}_T}}(k|t-1)$ should be set to 0. The corresponding unconstrained state estimate is referred to as $\hat{x}^{u*}_{\mathbf{j}_T}(k|t)$.

## 3.2 Constrained Filter Bank

The CFB will recompute the state estimates but now considering the constraints (6), (7) and (8). The constrained least-squares filter is somehow more complicated. First the least-squares state vector must be augmented to incorporate both the input disturbance and measurement noise vectors, since there exist explicit constraints on these variables:

$$\begin{bmatrix} x_{\mathbf{j}_T}(k)\\ W_{\mathbf{j}_T}(k)\\ V_{\mathbf{j}_T}(k)\end{bmatrix} \tag{14}$$

Notice that by explicitly considering the input disturbance and measurement noise sequences, all the uncertainty is removed from the observation equation (5) and it becomes an equality constraint:

$$\mathbf{H_e}\ .\begin{bmatrix} x_{\mathbf{j}_T}(k)\\ W_{\mathbf{j}_T}(k)\\ V_{\mathbf{j}_T}(k)\end{bmatrix} = \mathbf{h_e} \quad \Leftrightarrow \tag{15}$$

$$\Leftrightarrow\ \left[\mathbf{C}_{\mathbf{j}_T}\ \mathbf{L}_{\mathbf{j}_T}\ I_{n_Y}\right].\begin{bmatrix} x_{\mathbf{j}_T}(k)\\ W_{\mathbf{j}_T}(k)\\ V_{\mathbf{j}_T}(k)\end{bmatrix} = \left[Y_T(k)-\mathbf{D}_{\mathbf{j}_T}U_T(k)-\mathbf{g}_{\mathbf{j}_T}\right]$$

The constraints of the *dms* (6) and the bounds on the input disturbance and measurement noise vectors defined by the sets $\mathbb{W}_{\mathbf{j}_T}$ and $\mathbb{V}_{\mathbf{j}_T}$ described by equations (7) and (8) compose the inequality constraints of the least-squares problem, according to:

$$\mathbf{H_i}\ .\begin{bmatrix} x_{\mathbf{j}_T}(k)\\ W_{\mathbf{j}_T}(k)\\ V_{\mathbf{j}_T}(k)\end{bmatrix} \leq \mathbf{h_i} \quad \Leftrightarrow \tag{16}$$

$$\Leftrightarrow \begin{bmatrix}\mathbf{S}_{\mathbf{j}_T} & \mathbf{Q}_{\mathbf{j}_T} & 0\\ 0 & \mathbf{H}_{\mathbb{W}_{\mathbf{j}_T}} & 0\\ 0 & 0 & \mathbf{H}_{\mathbb{V}_{\mathbf{j}_T}}\end{bmatrix}.\begin{bmatrix} x_{\mathbf{j}_T}(k)\\ W_{\mathbf{j}_T}(k)\\ V_{\mathbf{j}_T}(k)\end{bmatrix} \leq \begin{bmatrix}\mathbf{T}_{\mathbf{j}_T}-\mathbf{R}_{\mathbf{j}_T}U_T(k)\\ \mathbf{h}_{\mathbb{W}_{\mathbf{j}_T}}\\ \mathbf{h}_{\mathbb{V}_{\mathbf{j}_T}}\end{bmatrix}$$

Having defined the constraints matrices, the constrained least-squares filter corresponding to the mode sequence $\mathbf{j}_T$ is given by:

$$\begin{bmatrix}\hat{x}_{\mathbf{j}_T}(k|t)\\ \hat{W}_{\mathbf{j}_T}(k|t)\\ \hat{V}_{\mathbf{j}_T}(k|t)\end{bmatrix} = \begin{bmatrix}\hat{x}_{\mathbf{j}_T}(z,k|t-1)\\ \hat{W}_{\mathbf{j}_T}(k|t-1)\\ \hat{V}_{\mathbf{j}_T}(k|t-1)\end{bmatrix} + \mathbf{K}_{\mathbf{j}_T}(k|t)\left(\begin{bmatrix}\mathbf{h_e}\\ \mathbf{h_i}\end{bmatrix}-\begin{bmatrix}\mathbf{H_e}\\ \mathbf{H_i}\end{bmatrix}.\begin{bmatrix}\hat{x}_{\mathbf{j}_T}(k|t-1)\\ \hat{W}_{\mathbf{j}_T}(k|t-1)\\ \hat{V}_{\mathbf{j}_T}(k|t-1)\end{bmatrix}\right) \tag{17}$$

The constrained least-squares filter gain is defined as:

$$\mathbf{K}_{\mathbf{j}_T}(k|t) = \tag{18}$$

$$\left(\begin{bmatrix}\Sigma_{x_{\mathbf{j}_T}}(k|t-1) & 0 & 0\\ 0 & \Sigma_{W_{\mathbf{j}_T}} & 0\\ 0 & 0 & \Sigma_{V_{\mathbf{j}_T}}\end{bmatrix}^{-1}+\begin{bmatrix}\mathbf{H_e}\\ \mathbf{H_i}\end{bmatrix}^{\mathrm{T}}\mathbf{Z}_{\mathbf{j}_T}(k|t)\begin{bmatrix}\mathbf{H_e}\\ \mathbf{H_i}\end{bmatrix}\right)^{-1}\begin{bmatrix}\mathbf{H_e}\\ \mathbf{H_i}\end{bmatrix}^{\mathrm{T}}\mathbf{Z}_{\mathbf{j}_T}(k|t)$$

where $\Sigma_{x_{\mathbf{j}_T}}(k|t-1)$ is the covariance matrix associated with the *a priori* state estimate $\hat{x}_{\mathbf{j}_T}(k|t-1)$. $\mathbf{Z}_{\mathbf{j}_T}(k|t)$ is the diagonal matrix that defines the active constraints.

There are several methods, most of them iterative, for determining the matrix $\mathbf{Z}_{\mathbf{j}_T}(k|t)$, or equivalently the set of active constraints. Here, the active set method presented in (Fletcher, 1987) will be used.

As in the unconstrained case, *a priori* information may be discarded by setting $\Sigma^{-1}_{x_{\mathbf{j}_T}}(k|t-1)$ to 0. The corresponding constrained state estimate is referred to as $\hat{x}^{c*}_{\mathbf{j}_T}(k|t)$.

## 3.3 Discrete Mode Sequence Estimator

The DMSE deals with the estimation of the discrete mode sequence and, consequently, selects the filter which will provide the final continuous state estimate.

According to the least-squares philosophy, an approximation of the measured output sequence is computed for every possible *dms* and then, the one providing the smallest squared error should be selected as the least-squares estimate.

The *dms* estimate is then selected as the one that presents the lowest constrained squared error, $\alpha^c_{\mathbf{j}_T}$:

$$\hat{\mathbf{i}}_T(k|t) = \arg\min_{\mathbf{j}_T} \alpha^c_{\mathbf{j}_T}(k|t) \tag{19}$$

The squared error associated with the *dms* $\mathbf{j}_T$ is given by:

$$\begin{aligned}\alpha_{\mathbf{j}_T}(k|t) &= \left\|\hat{Y}^*_{\mathbf{j}_T}(k|t) - Y_T(k)\right\|^2_{\Sigma^{-1}_{Y_{\mathbf{j}_T}}} = \\ &= \left[\hat{Y}^*_{\mathbf{j}_T}(k|t) - Y_T(k)\right]^{\mathrm{T}} \Sigma^{-1}_{Y_{\mathbf{j}_T}} \left[\hat{Y}^*_{\mathbf{j}_T}(k|t) - Y_T(k)\right]\end{aligned} \tag{20}$$

where:

$$\hat{Y}^*_{\mathbf{j}_T}(k|t) = \mathbf{C}_{\mathbf{j}_T}\hat{x}^*_{\mathbf{j}_T}(k|t) + \mathbf{D}_{\mathbf{j}_T}U_T(k) + \mathbf{g}_{\mathbf{j}_T} \tag{21}$$

and $\hat{x}^*_{\mathbf{j}_T}(k|t)$ is the estimated state of the *dms* $\mathbf{j}_T$ when all past information is discarded, $\left(\Sigma^{-1}_{x_{\mathbf{j}_T}}(k|t-1) = 0\right)$.

The squared errors computed by equation and (20) are useful when comparing continuous state estimates from the same *dms*. However, when the covariance matrices are different, an additional factor, $\bar{\alpha}_{\mathbf{j}_T}$, must be considered to allow a meaningful comparison between squared errors. Recalling the relation between least-squares and the maximization of the Gaussian likelihood function (or its logarithm), the value of $\bar{\alpha}_{\mathbf{j}_T}$ should be defined as:

$$\bar{\alpha}_{\mathbf{j}_T} = -\frac{1}{2}\ln\left((2\pi)^{n_Y}\det\left(\Sigma_{Y_{\mathbf{j}_T}}\right)\right) \tag{22}$$

Equation (20) should be modified to:

$$\alpha_{\mathbf{j}_T}(k|t) = \bar{\alpha}_{\mathbf{j}_T} + \left\|\hat{Y}^*_{\mathbf{j}_T}(k|t) - Y_T(k)\right\|^2_{\Sigma^{-1}_{Y_{\mathbf{j}_T}}} \tag{23}$$

Equation (23) can be used to compute the squared errors of both the unconstrained estimates, $\alpha^u_{\mathbf{j}_T}(k|t)$, and the constrained estimates, $\alpha^c_{\mathbf{j}_T}(k|t)$, using $\hat{x}^{u*}_{\mathbf{j}}(k|t)$ and $\hat{x}^{c*}_{\mathbf{j}}(k|t)$, respectively.

## 3.4 Computational Issues

Concerning computational requirements, it is noticed that there can be as many as $n_s^T$ *dms*, which becomes an extremely large number even for relatively small $n_s$ and $T$. So, computationally demanding calculations should be preformed for the minimum number of *dms* possible.

Analyzing the required computations one concludes that $\hat{x}^{u*}_{\mathbf{j}_T}(k|t)$ can be determined by simple matrix sums and multiplications if the filter gain $\mathbf{K}_{\mathbf{j}_T}(k|t-1)$ is computed off-line, since there are no varying terms as can be seen in equation (9). The corresponding squared error $\alpha^u_{\mathbf{j}_T}(k|t)$, computed through equation (23), can also be determined using simple matrix sums and multiplications from $\hat{x}^{u*}_{\mathbf{j}_T}(k|t)$. The continuous state estimate $\hat{x}^u_{\mathbf{j}_T}(k|t)$ on the other hand, requires a matrix inversion to determine the corresponding filter gain using equation (10) since the matrix $\Sigma^{-1}_{x_{\mathbf{j}_T}}(k|t-1)$ is not known in advance.

The constrained estimates require much more complex computations in the solution of the inequality constrained least-squares problem. An iterative algorithm has to be preformed online, and involves one matrix inversion at each iteration which is computationally heavy. There is the possibility that the solution corresponding to the true *dms* is the same as the unconstrained solution and the iterative algorithm stops at the first iteration. In general, however, this will not be the case. So, the computation of constrained solutions should only be done in cases of absolute necessity. The squared error of the constrained estimates $\alpha^c_{\mathbf{j}_T}(k|t)$ can be determined using simple matrix sums and multiplications from $\hat{x}^{c*}_{\mathbf{j}}(k|t)$.

The proposed algorithm should take these knowledge into account and arrive at the final estimates in the most efficient way possible.

To avoid the computation of the constrained least-squares estimates from all discrete mode sequences, the following relation between the constrained and unconstrained squared errors for a given discrete mode sequence is used:

$$\alpha^u_{\mathbf{j}_T}(k|t) \leq \alpha^c_{\mathbf{j}_T}(k|t) \tag{24}$$

An efficient reduction on the number of constrained estimates that have to be computed can be achieved by computing all unconstrained estimates $\hat{x}^{u*}_{\mathbf{j}}(k|t)$ and the corresponding squared errors $\alpha^u_{\mathbf{j}_T}(k|t)$ and then, start replacing the unconstrained solutions with the

corresponding constrained ones, from the lower values of the squared error. Whenever the lowest squared error corresponds to a constrained solution, the algorithm stops since no further reduction of the squared error can be done. The discrete mode sequence and continuous state estimates are the ones corresponding to that lowest squared error.

This algorithmic procedure may provide a substantial reduction in the number of inequality constrained least-squares problems to be solved since the increase in the squared error should be small, or even zero, for the true *dms*. However, the unconstrained solutions of incorrect *dms* may have low squared errors, which rise substantially only when the respective constrained solutions are computed. An efficient procedure to detect these incorrect *dms* before computing the respective constrained estimates would reduce the computational requirements even more.

To further improve the algorithm, the following $\mathcal{B}$ matrix must be introduced. Each coefficient $\beta_{\mathbf{i}_T,\mathbf{j}_T}$ of the matrix $\mathcal{B}$ is defined as the maximum value of $\alpha^c_{\mathbf{i}_T}$ under which $\alpha^c_{\mathbf{i}_T}$ is always smaller than $\alpha^c_{\mathbf{j}_T}$, or in an even more restrictive way, under which $\mathbf{j}_T$ is never the estimated sequence. The coefficients $\beta_{\mathbf{i}_T,\mathbf{j}_T}$ can be computed off-line by the following optimization problem, which falls in the general class of Second-Order Cone Programs for which efficient solvers have already been developed, for instance, by (Alizadeh and Goldfarb, 2001):

$$\begin{aligned} \beta_{\mathbf{i}_T,\mathbf{j}_T} = & \min_{Y_T,U_T} \ \alpha^c_{\mathbf{i}_T}(Y_T,U_T) \\ \text{subject to :} & \\ & U_T \in \mathbb{U}^T \\ & \hat{\mathbf{i}}_T = \mathbf{j}_T \end{aligned} \tag{25}$$

By this definition of $\beta_{\mathbf{i}_T,\mathbf{j}_T}$, when the constrained solution of a *dms* $\mathbf{i}_T$ is computed, all *dms* $\mathbf{j}_T$ such that $\beta_{\mathbf{i}_T,\mathbf{j}_T}$ is greater than $\alpha^c_{\mathbf{i}_T}(k|t)$ can be discarded. This algorithmic procedure provides an even greater reduction on the number of constrained problems to be solved. Notice that this procedure does not even require the computation of the unconstrained solutions of the *dms* to be discarded.

Both previous modifications to the algorithm require the existence of one constrained solution to discard any other *dms*. Furthermore, the number of discarded *dms* depends on the quality of the constrained solution. In the following, some attention will be given to the recursiveness of the DMSE and the methodology to determine the *dms* that will most likely provide good constrained estimates.

At a given time instant $t+1$ the following quantities have been computed at the previous time instant: the discrete mode sequence estimate, $\hat{\mathbf{i}}_T(k|t)$, the squared errors (or lower bounds) of all *dms*, $\alpha^c_{\mathbf{j}_T}(k|t)$ and, the continuous state estimates $\hat{x}^{c*}_{\mathbf{j}_T}(k|t)$ and the values of the estimated input disturbances $\hat{W}_{\mathbf{j}_T}(k|t)$ for the *dms* whose squared errors have been computed, including the *dms* estimate. These quantities allow the computation of the *a priori* continuous state estimate corresponding to the discrete mode sequence estimate at the following time instant:

$$\begin{aligned} \hat{x}^*_{\mathbf{j}_T}(t+1|t) = & \Big(A_{j(t)} \dots A_{j(k)}\Big)\hat{x}^*_{\mathbf{j}_T}(k|t) + \\ & \Big[A_{j(t)} \dots A_{j(k+1)}B_{j(k)}, \dots, B_{j(t)}\Big]U_T(k) + \\ & \Big[A_{j(t)} \dots A_{j(k+1)}W_{j(k)}\ , \dots,\ W_{j(t)}\Big]\hat{W}_{\mathbf{j}_T}(k|t) + \\ & \Big(A_{j(t)} \dots A_{j(k+1)}f_{j(k)} + \dots + f_{j(t)}\Big) \end{aligned} \tag{26}$$

This estimate can be used to obtain some insight on the likelihood of the discrete mode at the next time instant $j(t+1)$. The discrete modes $j(t+1)$ can be sorted by ascending values of:

$$\begin{aligned} & \gamma_{\mathbf{j}_T,j}(t+1|t) = \\ & \max\Big(S_j\,\hat{x}^*_{\mathbf{j}_T}(t+1|t) + R_j u(t+1) + Q_j \hat{w}(t+1|t) - T_j\Big) \end{aligned} \tag{27}$$

The value of $\hat{w}(t+1)$ should be set to $\mathrm{E}\{w_j\}$.

The discrete modes $j(t+1)$ that provide the lower values of $\gamma_{\mathbf{j}_T,j}(t+1|t)$ correspond the discrete mode sequences $\mathbf{j}_T = \big\{j(k+1), \dots, j(t), j(t+1)\big\}$ at time instant $t+1$ most likely to succeed to $\mathbf{j}_T$ at time instant $t$.

Applying this methodology to the discrete mode sequence estimate at the previous time instant, $\hat{\mathbf{i}}_T(k|t)$, should provide *dms* with very low squared errors that discard most of the other candidate *dms*. The same reasoning should be applied to all other discrete mode sequences of the previous time instant that have not been discarded yet, starting from the ones that present lowest squared errors and then the ones with the lowest bounds.

# 4 EXPERIMENTAL APPLICATION

To demonstrate the applicability of the hybrid estimation algorithms, the laboratory setup of the DTS200 three-tanks system from AMIRA® (Amira, 2002) will be used to simulate different situations common in hybrid estimation. A photo of the three-tanks system is presented in figure 2 showing the different components of the experimental setup. The plant consists of three plexiglas cylinders or tanks, $T_1$, $T_2$ and

$T_3$ with similar cross section. These are connected in series with each other by cylindrical pipes with cross section $S_n$. Located at $T_2$ is the single so called nominal outflow valve $V_0$ which also has a circular cross section $S_n$. The outflowing liquid (colored distilled water) is collected in a reservoir, which supplies the pumps $P_1$ and $P_2$. Here the water circuit is closed. $h_{max}$ denotes the highest possible liquid level in any of the tanks. In case the liquid level of $T_1$ or $T_2$ exceeds this limit the corresponding pump will be switched off automatically. $Q_1$ and $Q_2$ are the flow rates from pumps $P_1$ and $P_2$, respectively.

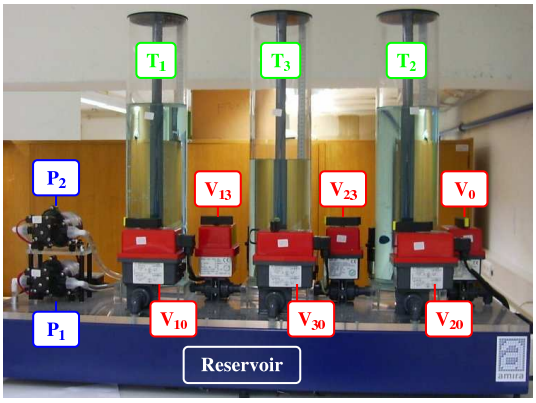



Figure 2: The three-tanks setup.

The pump flow rates $Q_1$ and $Q_2$ and the position of the valves $V_{13}$, $V_{23}$, $V_0$, $V_{10}$, $V_{20}$, $V_{30}$, denote the controllable variables, while the liquid levels of $h_1$, $h_2$ and $h_3$ are the output variables. The necessary level measurements are carried out by piezo-resistive difference pressure sensors. There are also potentiometric sensors that measure the position of each valve. The sensor signals are preprocessed to the interval $[0;1]$ and so need to be adjusted to $[0;h_{max}]$ for the water levels. For the remainder of this section the three-tanks system will be adapted so that more realistic hybrid estimation problems can be studied while simultaneously simplifying the presentation of results. The new model is present in figure 3 where the elements in grey are assumed to be nonexistent, the elements in green are fully operational and the elements in red may be subject to faults and will be used to model input disturbances.

Pump $P_1$ is considered to be a fully operational on/off valve. Valve $V_{13}$ will have two nominal values "on" and "off", while Valve $V_{10}$ will remain closed. Both these valves are subject to a possible fault resulting in an unmeasurable flow to cross them and described as an input disturbance. The water level sensor of tank 3 can also be subject to a fault. The Valve $V_{30}$ is considered to be a fully operational "on/off" valve with no possible faults, while Valves $V_{20}$, $V_{23}$ and $V_0$ will remain closed and so can be considered to be nonexistent.

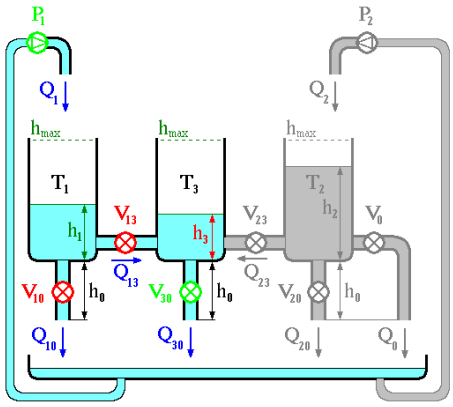



Figure 3: Final model of the three-tanks system.

The system can exhibit a large number of different dynamics, depending on the state of each discrete variable. The full hybrid model description of the system can be found in (Pina, 2007).

## 4.1 Estimation of the Fault in Valve $\mathbf{V_{10}}$

In this example, the estimation algorithm will have to estimate the discrete mode that indicates a fault on valve $\mathbf{V_{10}}$. As the analysis will focus on valve $\mathbf{V_{10}}$, the faults on valve $\mathbf{V_{13}}$ and sensor $\mathbf{h_3}$ will be considered nonexistent. A single test will be performed where various situations arise and are then analyzed separately. The system is excited according to the discrete variables presented in table 1. Various positions for the valve $\mathbf{V_{10}}$ are considered, corresponding to different intensities of the fault.

Table 1: Evolution of the discrete variables.

| Time(s) | 0-49 | 50-99 | 100-149 | 150-199 | 200-249 | 250-300 |
|---|---|---|---|---|---|---|
| $\mathbf{V_{10}}$ | "ok" | "faulty" *med* | "faulty" *max* | "faulty" *med* | "faulty" *max* | "ok" |
| $\mathbf{V_{13}}$ | "ok" | "ok" | "ok" | "ok" | "ok" | "ok" |
| $\mathbf{h_3}$ | "ok" | "ok" | "ok" | "ok" | "ok" | "ok" |
| $\mathbf{P_1}$ | "on" | "on" | "on" | "on" | "on" | "on" |
| $\mathbf{V_{13}}$ | "open" | "open" | "open" | "open" | "open" | "open" |
| $\mathbf{V_{30}}$ | "open" | "open" | "open" | "open" | "open" | "open" |

The measured outputs and the estimated water levels are presented in figure 4, where the influence of the intensity of the fault can be clearly seen.

The real (observed) and estimated values of the fault using the IMM algorithm are shown in figure 5. As the fault in valve $\mathbf{V_{10}}$ takes one time instant to be reflected in the water level measurements, only the value of $f_{V_{10}}(k-1|k)$ is relevant. Note that $f_{V_{10}}(k-1|k)$ is a discrete variable that takes value 1 when a leak occurs, and value 0 when there is no fault.

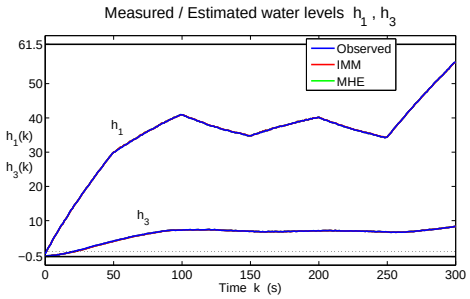



Figure 4: Water levels estimation using the IMM and MHE estimation algorithms.

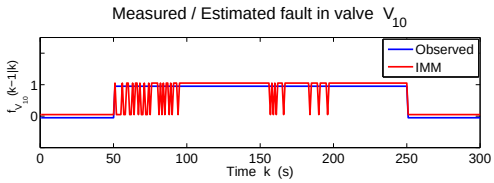



Figure 5: Estimation of the discrete mode sequence relative to the fault in valve $\mathbf{V_{10}}$.

The corresponding estimated continuous input disturbances by both algorithms are shown in figure 6. As the fault in valve $\mathbf{V_{10}}$ takes one time instant to be reflected in the water level measurements, only the value of $w_{V_{10}}(k-1|k)$ is estimated. The variable $w_{V_{10}}$ determines the leaking flow and is considered to be a uniformly distributed random variable defined in the interval $[-0.4;0.4]$ cm, with zero mean and variance $\frac{0.8^2}{12}$ cm$^2$ for all $k$, where 0.8 is the maximum water level change when the valve $\mathbf{V_{10}}$ is fully open.

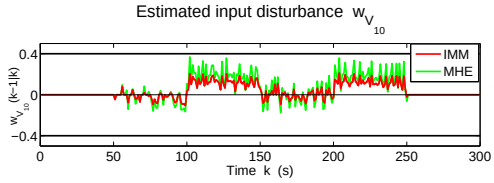



Figure 6: Estimation of the input disturbance $w_{V_{10}}(k-1|k)$ corresponding to the fault in valve $\mathbf{V_{10}}$.

The difference observed in both algorithms for the estimation of the disturbance $w_{V_{10}}(k-1|k)$ shows that the MHE algorithm is not able to weight the disturbance with any prior value so allowing it to change freely, which increases the variation of the input disturbance estimates.

The estimation results presented in figures 4 and 5 will now be analyzed independently for the 3 considered valve $\mathbf{V_{10}}$ fault intensities.

### 4.1.1 Case 1 - Fault Inactive

For time intervals $[0;50[$ s and $[250;300]$ s valve $\mathbf{V_{10}}$ remained closed and the fault is considered inactive. Despite being inactive, there is still a possibility of a wrong estimate reflected on the value of the discrete variable $f_{V_{10}}$. However, as shown in figure 7, the valve's true state was correctly estimated during these time periods.

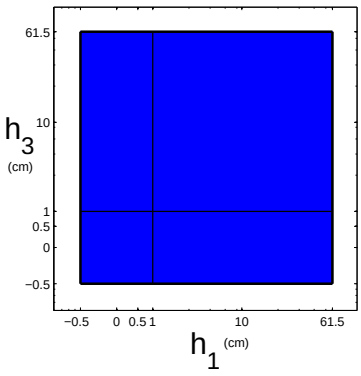



Figure 7: Map of probability of correct mode estimation, with 1s delay, when valve $\mathbf{V_{10}}$ is fully closed. (Red - probability of correct mode estimation 0, Blue - probability of correct mode estimation 1)

Figure 7 shows that if the valve $\mathbf{V_{10}}$ is closed there is no possibility of estimating a discrete mode sequence corresponding to an open valve condition. Thus the inactive fault is always correctly estimated.

### 4.1.2 Case 2 - Fault Active with Intermediate Intensity

The valve $\mathbf{V_{10}}$ has an intermediate open position during time intervals $[50;100[$ s and $[150;200[$ s allowing an unmeasured flow to cross it. In this case, a fully closed valve was estimated by the IMM algorithm in several time instants. These wrong estimates are understandable since the effect on the water level of tank 1 is not too drastic and can be mistaken by any other source of uncertainty, like measurement noise for instance. This difficulty in discerning whether the valve is slightly open or fully closed is patent in the map of probability of correct mode estimation shown in figure 8. It can also be concluded that the probability of an incorrect estimation of the valve's condition increases as the water level of tank 3 becomes lower.

The map of probability of correct mode estimation is not able to show the existing dependence between the probability of correctly determining the valve's condition and its real position. It is clear from figure

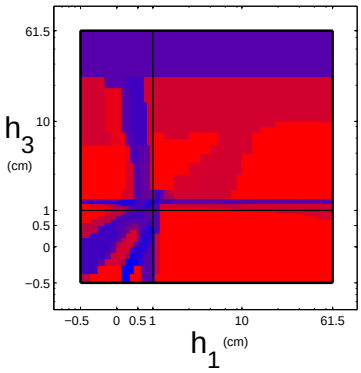



Figure 8: Map of probability of correct mode estimation, with 1s delay, when fault $f_{V_{10}}$ is active. (Red - probability of correct mode estimation 0, Blue - probability of correct mode estimation 1)

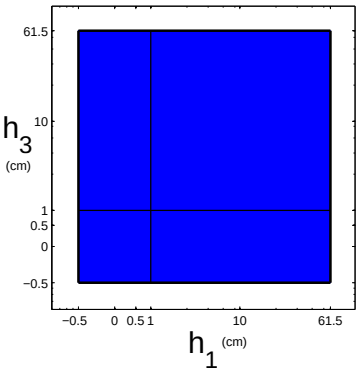



Figure 9: Map of probability of correct mode estimation, with 1s delay, considering that fault $f_{V_{10}}$ has maximum intensity, $w_{V_{10}} = 0.4$. (Red - probability of correct mode estimation 0, Blue - probability of correct mode estimation 1)

4 that the valve $\mathbf{V_{10}}$ is more closed during the time interval $[50\,;100[$s than in $[150\,;200[$s. This fact is reflected in a higher number of incorrect mode sequence estimations in case the valve remains closer to its nominal closed position. The following case will further explore this dependence.

#### 4.1.3 Case 3 - Fault Active with Maximum Intensity

If valve $\mathbf{V_{10}}$ is fully open it becomes much easier to determine its position, thus allowing the IMM algorithm to provide correct estimates for the discrete mode sequence during time intervals $[100\,;150[$s and $[200\,;250[$s. This is quite obvious since the effect on the water level of tank 1 is very intense and can not be mistaken by any other source of uncertainty. This result is depicted in figure 9.

This map of probability of correct mode estimation was computed considering an hypothetical model for the system where valve $\mathbf{V_{10}}$ can only be fully open or fully closed.

Figure 9 shows that when the fault $f_{V_{10}}$ has maximum intensity, $w_{V_{10}} = 0.4$, it is always correctly estimated. However, further results have shown that for very low water levels in tank 1 the difference between a fully open or fully closed valve are reduced, being even undetectable when the tank is empty. This is explained by the fact that the maximum fault intensity allowed by the model, $w_{V_{10}} = 0.4$, can not be achieved in practice when tank 1 is almost empty but rather when it is full.

## 5 CONCLUSIONS

This paper presented an efficient hybrid estimation algorithm based on an IMM setup composed by a set of least-squares filters. The computational efficiency is obtained by some algorithmic procedures that discard many candidate *dms* before performing heavy computations. These procedures rely on the early determination of good estimates, on the separation of constrained and unconstrained estimates and on some bounding parameters for the squared errors.

The IMM was able to provide accurate online estimates for both continuous states and discrete variables when applied to the hybrid model of the benchmark AMIRA DTS200 three-tanks system experimental setup. The potential of the IMM algorithm was demonstrated when comparing its computational efficiency with the MHE with unknown inputs algorithm for a fault detection problem.

One of the most relevant issues that influence the computational efficiency of hybrid methodologies has to do with the high number of discrete modes that are tipically involved in a medium size hybrid system model. This fact eventually turns most of the problems untractable. For the case of the three-tanks system experimental setup, it was noticed that the consideration of all three tanks in the same hybrid model requires huge computational resources. Thus, authors believe that a multi-agent modeling architecture can significantly simplify the all model complexity while being able to retain its full hybrid dynamical flavour. As the size of the problems to be solved with hybrid systems grows exponentially with the number of discrete modes involved, multi-agent architectures may be the solution to the huge complexity of hybrid

methodologies, thus being a very interesting and possibly fruitful research topic.

## REFERENCES


Alessandri, A. and Coletta, P. (2003). Design of observers for switched discrete-time linear systems. In *Proc. American Control Conference*, pages 2785–2790, Denver, Colorado.

Alizadeh, F. and Goldfarb, D. (2001). Second-order cone programming. Technical Report RRR Report number 51-2001, RUTCOR, Rutgers University, Piscataway, New Jersey.

Amira (2002). *DTS200 - Laboratory Setup Three-tank-system*. Amira, Duisburg, Germany.

Antsaklis, P. (2000). A brief introduction to the theory and applications of hybrid systems. *Proc. IEEE, Special Issue on Hybrid Systems: Theory and Applications*, 88(7):879–886.

Athans, M. and Chang, C. (1976). Adaptive estimation and parameter identification using multiple model estimation algorithm. Technical Report 28, M.I.T. - Lincon Laboratory, Lexington, Massachusetts.

Balluchi, A., Benvenuti, L., Benedetto, M. D., and Sangiovanni-Vincentelli, A. (2002). Design of observers for hybrid systems. In *Hybrid Systems: Computation and Control*, volume 2289 of *Lecture Notes in Computer Science*, pages 76–89. Springer Verlag.

Bemporad, A. and Morari, M. (1999). Control of systems integrating logic, dynamics, and constraints. *Automatica*, 35(3):407–427.

Blom, H. A. P. and Bar-Shalom, Y. (1988). The interactive multiple model algorithm for systems with markovian switching coefficients. *IEEE Trans. on Automatic Control*, 33(8):780–783.

Böker, G. and Lunze, J. (2002). Stability and performance of switching Kalman filters. *International Journal of Control*, 75(16/17):1269–1281.

Ferrari-Trecate, G., Mignone, D., and Morari, M. (2002). Moving horizon estimation for hybrid systems. *IEEE Trans. on Automatic Control*, 47(10):1663–1676.

Fletcher, R. (1987). *Practical methods of optimization*. A Wiley Interscience Publication, Chichester, New York, $2^{nd}$ edition.

Heemels, W., Schutter, B. D., and Bemporad, A. (2001). Equivalence of hybrid dynamical models. *Automatica*, 37(7):1085–1091.

Kamen, E. (1992). Study of linear time-varying discrete-time systems in terms of time-compressed models. In *Proc. 31th IEEE Conf. on Decision and Control*, pages 3070–3075, Tucson, Arizona.

Mazor, E., Averbuch, A., Bar-Shalom, Y., and Dayan, J. (1998). Interacting multiple model methods in target tracking: A survey. *IEEE Trans. on Aerospace and Electronic Systems*, 34(1):103–123.

Pina, L. (2007). *Hybrid state estimation*. PhD thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa, Portugal.

Pina, L. and Botto, M. A. (2006). Simultaneous state and input estimation of hybrid systems with unknown inputs. *Automatica*, 42(5):755–762.

Sontag, E. (1981). Nonlinear regulation: The piecewise linear approach. *IEEE Trans. on Automatic Control*, 26(2):346–358.

Torrisi, F. and Bemporad, A. (2001). Discrete-time hybrid modeling and verification. In *Proc. 40th IEEE Conf. on Decision and Control*, pages 2899–2904, Orlando, Florida.


## BRIEF BIOGRAPHY

Miguel Ayala Botto received the master degree in Mechanical Engineering in 1992 and the Ph.D. in Mechanical Engineering in 1996 from Instituto Superior Tcnico, Tecnhical University of Lisbon, Portugal. He spent the year of 1995 at the Control Laboratory, Department of Electrical Engineering, Delft University of Technology, Holland. Further, in the winter semester of the academic year 1999/2000 he held a postdoctoral position at the same laboratory. Since 2001 he is Associate Professor at the Department of Mechanical Engineering, Instituto Superior Tcnico, Portugal. He is currently coordinator of the research group on Systems and Control from the Center of Intelligent Systems of IDMEC - Institute of Mechanical Engineering. Since 2005 he is the head of the Portuguese Association on Automatic Control, the National Member Organization from IFAC. He has published more than 70 journal papers, book chapters, and communications in international conferences. He has been awarded in 1999 with ”The Heaviside Premium”, attributed by the Council IEE - The Institution of Electrical Engineers, UK. Currently he is Associate Editor of the International Journal of Systems Science (Taylor & Francis) and member of the IFAC Technical Committee on Discrete Event and Hybrid Systems. His main research interest is in the field of estimation and control of hybrid dynamical systems.

# DISTRIBUTED TECHNOLOGY FOR GLOBAL DOMINANCE

Peter Simon Sapaty
*Institute of Mathematical Machines and Systems, National Academy of Sciences*
*Glushkova Ave 42, 03187 Kiev, Ukraine*
*Tel: +380-44-5265023, Fax: +380-44-5266457*
*sapaty@immsp.kiev.ua*

Keywords: Global dominance, spatial scenarios, world processing language, distributed interpretation, emergency management, sensor networks, directed energy systems, avionics, electronic warfare, distributed objects tracking, collective behavior.

Abstract: A flexible, ubiquitous, and universal solution for management of distributed dynamic systems will be presented. It allows us to grasp complex systems on a higher than usual, semantic level, penetrating their infrastructures, also creating and modifying them, while establishing local and global dominance over the system organizations and coordinating their behavior in the way needed. The approach may allow the systems to maintain high runtime integrity and automatically recover from indiscriminate damages, preserving global goal orientation and situation awareness in unpredictable and hostile environments.

## 1 INTRODUCTION

We are witnessing a rapid growth of world dynamics caused by consequences of global warming, globalization of economy, numerous ethnic, religious and military conflicts, and international terrorism. To match this dynamics and withstand numerous threats and possible adversaries, effective integration of any available human and technical resources is crucial. These resources may be scattered and emergent, lacking the infrastructures and authorities for organization of the solutions needed, in real time and ahead of it.

Just communication between predetermined parts and systems with possible sharing a common vision, often called "interoperability", may not be sufficient. The whole distributed system (or system of systems) should rather represent a highly dynamic and integral organism, in which parts may be defined and interlinked dynamically in subordination to the global organization and system goals, which can vary at runtime, with the coined term "overoperability" (Sapaty, 2002) becoming more appropriate.

A related ideology and accompanying information & control technology, allowing us to provide a much higher than usual level of system understanding and control, will be outlined in this paper.

## 2 THE WORLD PROCESSING PARADIGM

Within the approach developed, a network of intelligent modules (U, see in Fig. 1), embedded into important system points, collectively interprets mission scenarios in a special high-level language, which can start from any nodes, covering the networked systems at runtime.

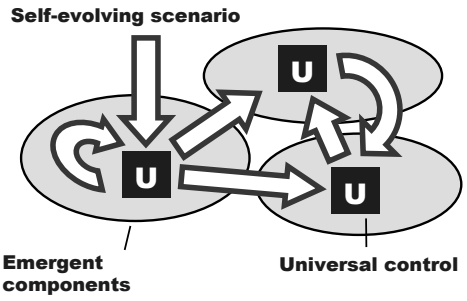



Figure 1: Runtime coverage of a distributed system.

The system "conquering" scenarios are integral and compact, being often capable of self-recovery after damages. They may be created on the fly, as traditional synchronization, data, code, and agents handling and exchanges are effectively shifted to the automatic implementation. This (parallel and fully

distributed, without central resources) spatial process can take into account details of the environments, which may be unpredictable and hostile, in which mission scenarios evolve.

Initially represented in a unified and compact form, the scenario and resources which may be needed for its development, can start from any system point (as shown in Fig.2).

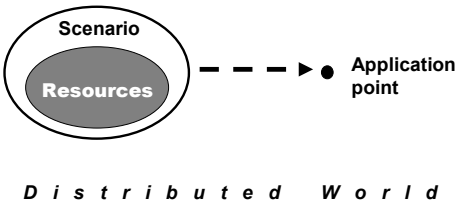



Figure 2: The initial state.

The scenarios can self-split, replicate, and modify while covering the distributed world or its part(s) needed at runtime, bringing operations and (both virtual and physical) resources into different points, also lifting, activating, and spreading further other scenarios and resources, already accumulated in the navigated world, as in Fig. 3.

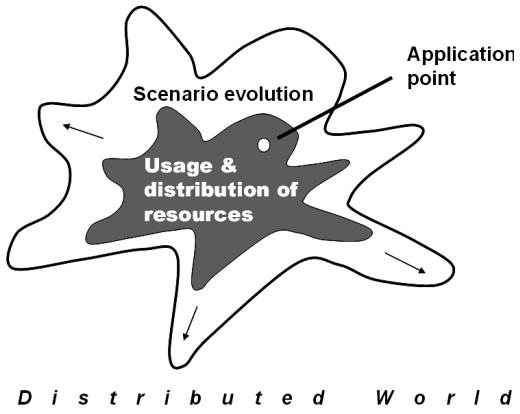



Figure3: Spreading operations and resources.

This is causing movement of information and physical matter, as well initiating interactions between manned and unmanned components, command and control (C2) including, as in Fig. 4 (S is for spatial scenarios or their parts, and R – for resources to implement the scenarios).

The main difference of this approach with the other works is that it describes on a higher level, in a concise way, of what the system should do or how should behave as a whole, while delegating numerous routines of partitioning into components (agents), with their interaction and synchronization, to the effective automatic level, while other approaches used to do the latter manually, and from the start. The approach can, however, describe and implement the system organization and its behavior at any levels needed, which may include:

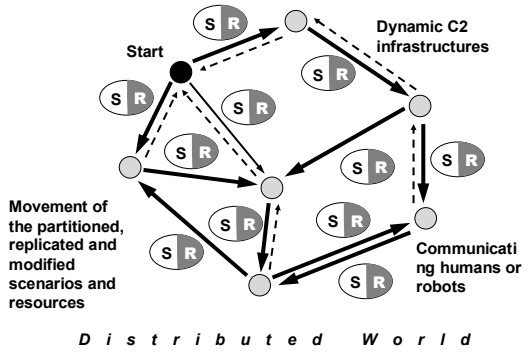



Figure 4: Resultant interactions between system parts.

- Most general, semantic, task formulation.
- Explicit projecting intelligence, information, matter, and power into particular physical or virtual locations, with doing jobs directly in the places reached, and if needed, cooperatively.
- Creating new active physical, virtual, or combined worlds, and organizing & coordinating their activity.
- Setting up implementation details, at any levels, say, for optimization of the use of scarce resources.

# 3 THE WORLD PROCESSING LANGUAGE (WPL)

This ideology and technology are based on the World Processing Language, WPL (Sapaty, 2005) describing what to do in distributed spaces rather than how to do, and by which resources (or even system organization), leaving these to the automatic interpretation in networked environments. The WPL fundamentals include:

- Association of any action with a position in physical, virtual, or combined space.
- Working with both information and physical matter.
- Runtime creation of distributed knowledge networks.
- Unlimited parallelism.
- Free movement or navigation in physical, virtual, or combined worlds.
- Fully distributed decision making with high integrity as a whole.
- Automatic command and control.

It is a higher-level language to efficiently command and control emergent human teams and armies. It is also a fully formal language suitable for automatic interpretation by mobile robots and their groups. Due to peculiar syntax and semantics, its parallel interpretation in distributed systems is straightforward, transparent, and does not need any central resources. Such complex problems as synchronization of multiple activities and collective (swarm as well as centrally or hierarchically controlled) behavior can be solved automatically by the networked interpreter, without traditional load on human managers and programmers.

This dramatically simplifies application programming, which is often hundreds of times more concise (and simpler) than in traditional programming languages. WPL allows for a direct access to the distributed world, performing any operations in any its points over local or remote data, which may represent both information and physical matter. Navigating in the world, WPL can modify it or even create from scratch, if required. Different movements and operations can be performed simultaneously and in parallel, and these may be free or may depend on each other.

WPL has a recursive syntax which can be expressed on the top level as follows (square brackets are for an optional construct, braces mean construct repetition with a delimiter at the right, and vertical bar separates alternatives).

***wave*** → ***constant*** | ***variable*** | [ ***rule*** ] ( {***wave*** , } )
***constant*** → ***information*** | ***matter***
***variable*** → ***nodal*** | ***frontal*** | ***environmental***
***rule*** → ***evolution*** | ***fusion*** | ***verification*** | ***essence***
***evolution*** → ***expansion*** | ***branching*** | ***advancing*** | ***repetition*** | ***granting***
***fusion*** → ***echoing*** | ***processing*** | ***constructing*** | ***assignment***
***verification*** → ***comparison*** | ***membership*** | ***linkage***
***essence*** → ***type*** | ***usage***

A rule is a very general construct, which, for example, can be:

- Elementary arithmetic, string or logic operation.
- Hop in physical, virtual, or combined space.
- Hierarchical fusion and return of (remote) data.
- Parallel and distributed control.
- Special context for navigation in space.
- Sense of a value for its proper interpretation.

Different types of variables, especially when used together, allow us to create efficient spatial algorithms which work "in between components" of distributed systems rather than in them. The variables called *nodal* can store and access local results in the system points visited, while others ones can move data in space together with the evolving control (*frontal variables*) or can access and impact the world navigated (*environmental variables*).

# 4 ELEMENTARY EXAMPLES

## 4.1 Setting Global Dominance

Let us assume that a node in the distributed system (see Fig.5) wants to establish the field of its dominance over other nodes which have a lower rank that itself (here the content, or name, of each node is considered as its rank).

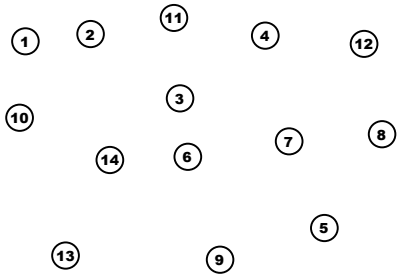



Figure 5: Distributed system nodes.

The following parallel and distributed program, applied in this node, spreads its own rank throughout the whole system in the frontal variable Rank. This puts the rank into the nodal variable Dominance in each visited node, if Rank exceeds the already existing value in Dominance (by the first access, the variable Dominance is assigned the value of the personal rank of each node).

```
frontal Rank = CONTENT;
nodal Dominance;
repeat (
  if (Dominance == nil, Dominance = CONTENT);
  if (Dominance < Rank,
      (Dominance = Rank; hop all neighbors),
        stop))
```

If applied, say, in node 11, this distributed program establishes only a partial dominance in the system, as shown in Fig. 6.

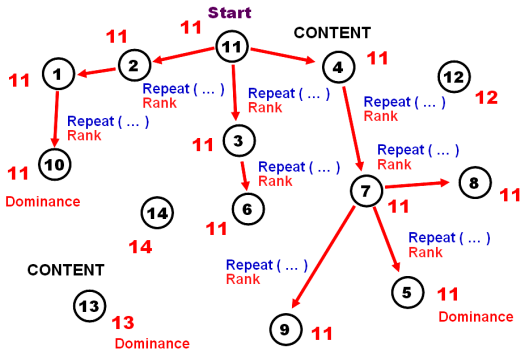



Figure 6: Resulting in partial dominance.

That will not be the case for node 14, which will set up its absolute dominance over the whole world by the program above, as in Fig. 7.

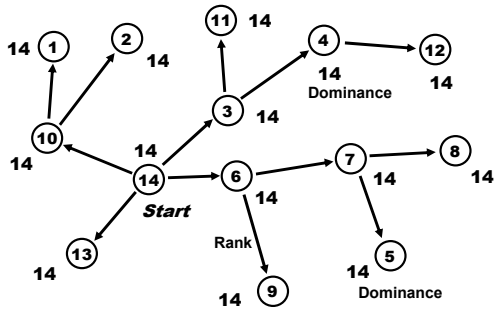



Figure 7: Setting absolute dominance of node 14.

## 4.2 Creating Infrastructures in the Distributed Space

It is easy to set up any infrastructures in the distributed space by the approach presented, with any topology. The following program, starting from node 3, will create (in parallel and distributed way) the networked structure shown in Fig. 8 over the set of already existing nodes.

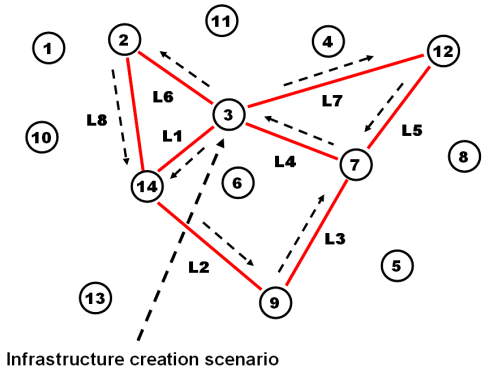



Figure 8: Creating a distributed infrastructure.

```
hop node 3;
create links ((L6# 2; L8#14),
  (L7#12; L5#7; L4#3), (L1#14; L2#9; L3#7))
```

Any functionality can be associated with both nodes and links of the obtained infrastructure at runtime, which will be operating as a system for the purpose needed.

## 4.3 Finding Patterns in the Infrastructure

It is convenient to find any patterns in the distributed infrastructures in WPL. Let such a pattern be a triangle, and we would like to find all of them in the infrastructure created. The following spatial program, starting in any node, does this, with listing resultant nodes of the triangles in their descending ranks.

```
hop all nodes; frontal (Triangle) = CONTENT;
twice (hop all links; CONTENT < BACK;
       Triangle &= CONTENT);
hop all links; element (Triangle, first) == CONTENT;
output Triangle
```

The result, issued in the node where the program was injected, will be as: (14, 3, 2), (12, 7, 3) -- see Fig.9.

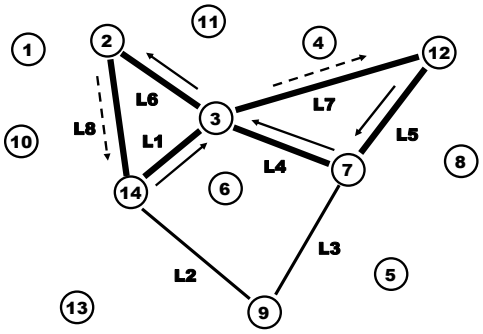



Figure 9: Finding all triangles in the infrastructure.

# 5 WPL INTERPRETER

The WPL interpreters may be embedded in internet hosts, robots, mobile phones, or smart sensors (an interpreter can also be a human being herself, understanding and executing high-level orders in WPL, while communicating with other humans or robots via WPL too). The interpreters may be concealed, if needed (say, to work in a hostile system); they can also migrate freely, collectively executing (also mobile) mission scenarios, resulting

altogether in the extremely flexible and ubiquitous system organization.

The basic WPL interpreter organization (Sapaty, 1993, 1999, 2005) is shown in Fig, 10, which may have both software and hardware implementation (the latter as "wave chip").

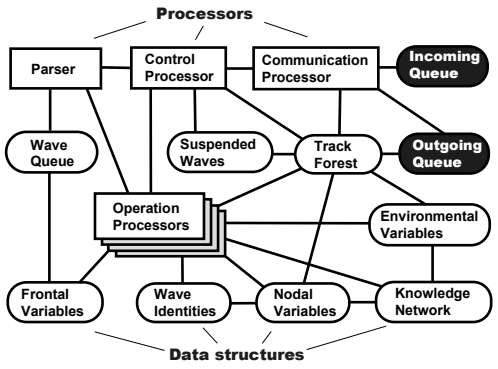



Figure 10: The WPL interpreter architecture.

The interpreter 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 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 interpreters can be embedded into any other systems, like mobile robots, allowing them to behave as integral teams, as shown in Fig. 11.

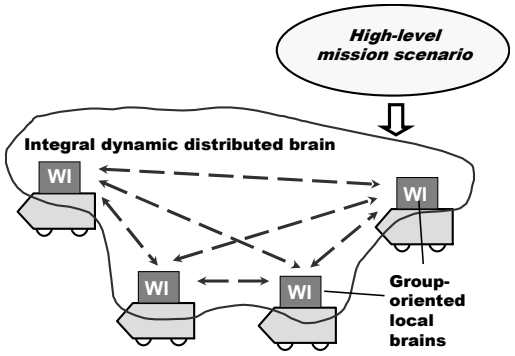



Figure 11: WPL interpreters (WI) forming distributed robotic brain.

# 6 EMERGENCY MANAGEMENT

Emergency management, EM (Sapaty, Sugisaka, Finkelstein, et al., 2006), due to the increased world dynamics, is becoming one of the hottest topics today. The emergency managers around the world are faced with new threats, new responsibilities, and new opportunities. Novel technologies, like the one of this paper, can alleviate consequences of natural (say, due to global warming) or manmade (like war conflicts) disasters. They can allow law enforcement and intelligence investigators to identify potential terrorist plots and then mount preemptive strikes to stop their plans.

The technology described can help in solving many EM problems by using communicating interpreters embedded in different electronic devices like, for example, laptops or mobile phones, with some disaster situation shown in Fig 12.

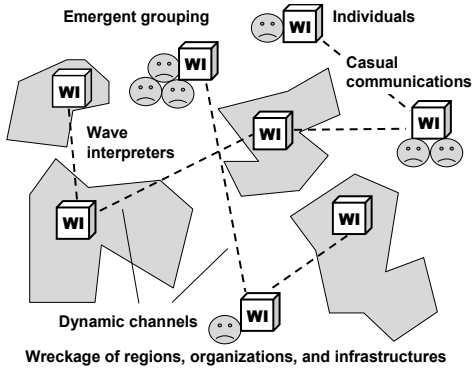



Figure 12: A disaster area with WPL interpreters embedded.

A very simple example may be here as a necessity to count the total number of casualties in the disaster area, on all its affected regions.

The following program can be applied from any WI as an entry one, which can reside within the disaster area or be away from it, and then can self-spread via local communications, organizing the whole region with embedded interpreters to work as an integral spatial supercomputer.

```
frontal Area = <disaster area definition>;
output sum (
  hop (directly, first come, nodes(Area));
  repeat(
    done(count casualties),
    hop(any links, first come, nodes(Area))))
```

More complex operations which can be organized in WPL may include the delivery of relief aid, an organized evacuation from the disaster area, and

organization of and cooperation with the rescue teams (which may include robotic components).

# 7 SENSOR NETWORKS

Sensor networks are a sensing, computing and communication infrastructure that allows us to instrument, observe, and respond to phenomena in the natural environment, and in our physical and cyber infrastructure. The sensors themselves can range from small passive microsensors to larger scale, controllable platforms. Typical applications of wireless sensor networks (WSN) include monitoring, tracking, and controlling. Some of the specific applications are habitat monitoring, object tracking, nuclear reactor controlling, fire detection, traffic monitoring, etc. Any distributed problems can be solved by dynamic self-organized sensor networks working in WPL (Sapaty, 2007a).

Starting from all transmitter nodes, the following program regularly (with interval of 20 sec.) covers stepwise, through local communications between sensors, the whole sensor network with a spanning forest, lifting information about observable events in each node reached, as shown in Fig. 13. Through this forest, by the internal interpretation infrastructure, the data lifted in nodes is moved and fused upwards the spanning trees, with final results collected in transmitter nodes and subsequently sent outside the system in parallel.

```
hop (all transmitters);
loop (
 sleep (20);
 IDENTITY = TIME;
 transmit (
  fuse (
     repeat (free (observe (events));
     hop (directly reachable, first come)))))
```

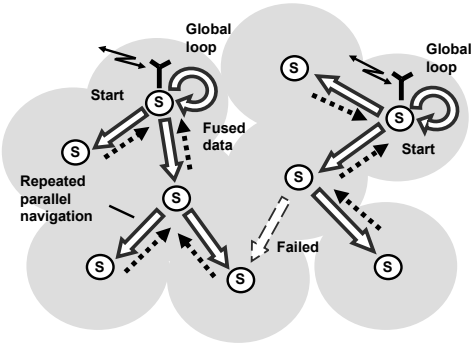



Figure 13: Collecting data by a sensor network.

Another program, below, provides for spanning tree coverage of some distributed phenomenon, with hierarchical collection, merging and fusing partial results got from different sensors into the global picture. The latter will be forwarded to a nearest transmitter via the previously created infrastructure with links infra, as shown in Fig. 14.

```
hop (random, all nodes, detected phenomenon).
loop (
  frontal Full = fuse (
    repeat (
      free (collect phenomenon),
      hop (directly reachable, first come,
            detected phenomenon)));
   repeat (hop links (-infra)). Transmit Full)
```

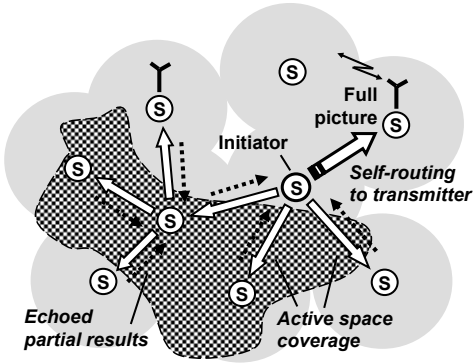



Figure 14: Space coverage with hierarchical assembling of a distributed phenomenon.

In more complex situations, which can be effectively programmed in WPL too, we may have a number of simultaneously existing phenomena, which can intersect in a distributed space. We may also face a combined phenomenon integrating features of different ones. The phenomena (like flocks of birds, manned or unmanned groups or armies, spreading fire or flooding) covering certain regions may change in size and shape, they may also move as a whole, preserving internal organization. All these situations can be managed in WPL.

# 8 DIRECTED ENERGY SYSTEMS

Directed energy (DE) systems are of a growing interest for broad applications in the nearest future, especially in infrastructure protection and defense. The DE-based systems will be able to operate under flexible command and control in WPL, restructuring and recovering in unpredictable environments without loss of functionality (Sapaty, Morozov, Sugisaka, 2007).

An elementary DE-based system may consist of a control center, DE source, relay mirror (RM), and target. Using WPL, the system functionality can be set up dynamically, on the fly, as by the following program:

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

Three snapshots of the system operation under this program are shown in Figs. 15-17.

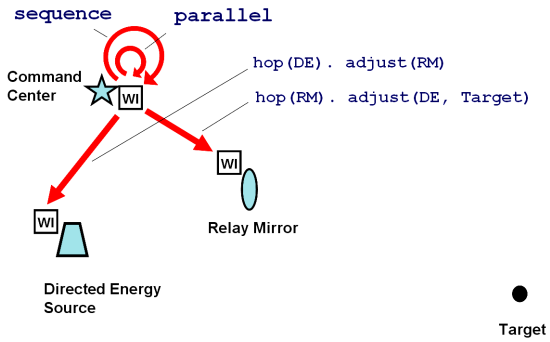



Figure 15: DE system operation, Snapshot 1.

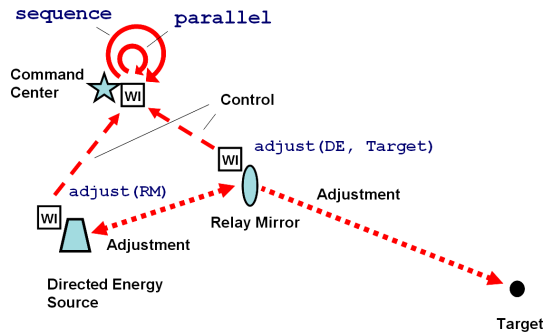



Figure 16: DE system operation, Snapshot 2.

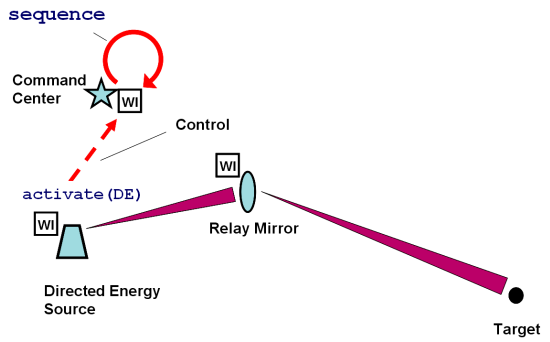



Figure 17: DE system operation, Snapshot 3.

Boeing's Advanced Relay Mirror System (ARMS) concept plans to entail a constellation of as many as two dozen orbiting mirrors that would allow a constant coverage of every corner of the globe. When activated, this would enable a directed energy response to critical trouble spots anywhere.

We will show here, be the program below, how the shortest path tree (SPT) starting from any DE source and covering the whole set of distributed mirrors can be created at runtime with the use of the technology presented. This will enable us to make optimal delivery of the directed energy to any point of the globe. The distributed SPT creation process is shown in Fig. 18.

```
nodal (Distance, Predecessor);
frontal (Length, Range = 400);
hop (DE);
Distance = 0. Length = 0;
repeat (
  hop (Range, all);
  Length += between (WHERE, BACK);
  or (Distance == nil, Distance > Length);
  Distance = Length; Predecessor = BACK)
```

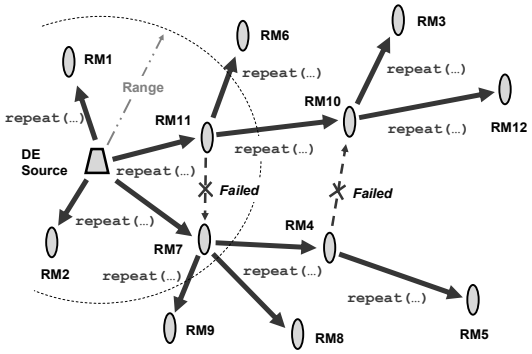



Figure 18: Dynamic shortest path tree through all RMs.

In case the target is defined, the following program forms a path from the DE source to the target via the relay mirrors, using the SPT formed, with a subsequent activation of the DE source to impact the target, as depicted in Fig. 19.

```
adjust (Seen (range), Predecessor);
repeat (
  hop (Predecessor, first);
  adjust (BACK, Predecessor));
activate (DE)
```

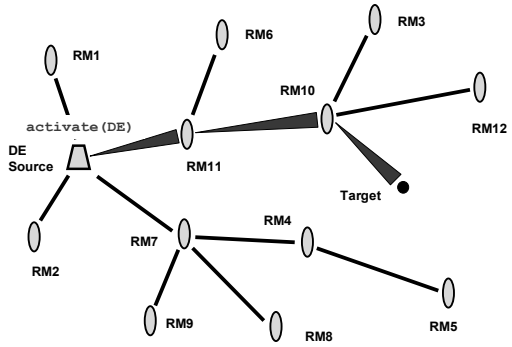



Figure 19: Energy delivery via the path found.

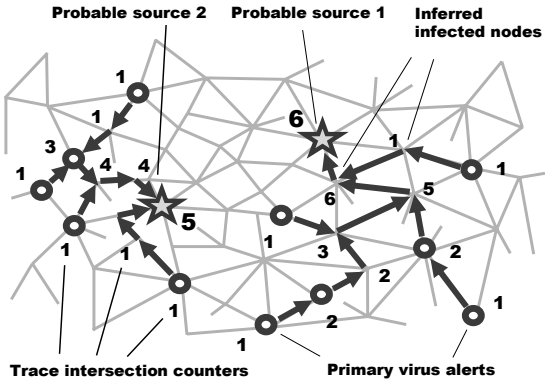



Figure 20: Finding virus sources in parallel.

# 9 ELECTRONIC WARFARE

Electronic warfare (EW) is becoming one of the main technological challenges of this century. All existing and being developed electronic support, attack, and protection measures usually have a very limited scope and effect if used alone. But taken together they may provide a capability for fulfilling the rapidly growing needs. Traditional communication and cooperation between these systems may not be sufficient. They should comprise altogether a much more integral system of systems with global situation awareness and "global will", which can be expressed and provided in WPL (Sapaty, 2007).

One of the typical EW tasks is fighting malicious intrusions and viruses in computer networks. Being itself a super-virus on the implementation level, the technology proposed, via the embedded network of WPL interpreters, can simultaneously discover and analyze electronic viruses, with blocking their spread and inferring attack sources. For example, the following scenario can find all virus sources in parallel, as shown in Fig 20:

```
nodal (Trace, Predecessor);
sequence (
   (hop (all nodes);
   nonempty (check general (viruses));
   repeat (
     increment (Trace);
     nonempty (Predecessor = check special
(viruses));
     hop (Predecessor))),
 output (
  sort (
    hop (all nodes); empty (Predecessor);
    nonempty (Trace); Trace & ADDRESS)))
```

# 10 AVIONICS

Avionics, or aviation electronics, represents a substantial share of the cost of any modern flying devices.

- Any avionics system, whether for a single aircraft or a group of them with manned or unmanned units, may be considered as a complex organization consisting of numerous components properly interacting with each other to pursue global goals. This organization can be effectively expressed in WPL on a variety of levels.
- This organization can be made flexible enough to recover from indiscriminate damages and restructure at runtime.
- The WP approach may offer real possibilities for a runtime recovery after damages, including reassembling of the whole system (or what remains of it) from any point.

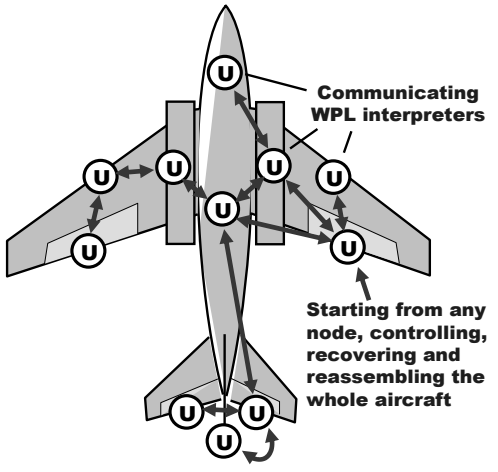



Figure 21: Aircraft self-analysis by the WPL network.

Implanting communicating WPL interpreters into main components of an aircraft, as universal control modules U (see Fig. 21), may allow us to convert the whole distributed object into a parallel computer capable of solving a variety of complex problems at runtime, including aircraft's safety and recovery (Sapaty, 2008).

The following program, starting from any point, is collecting availability of vital mechanisms of a damaged aircraft, analyzing their completeness to operate as a system, with making proper decisions (which may include the alarm with emergent evacuation of the crew).

```
nodal Available_Set =
   repeat (
      free (if CONTENT belongs_to
               (left_aileron, right_aileron, left_elevator,
                right_elevator, rudder, left_engine,
                right_engine,left_chassis,
right_chassis, …)
                  then CONTENT),
      hop_first  all_neighbors);
if sufficient Available_Set
      then control_with Available_Set
           otherwise alarm
```

# 11 DISTRIBUTED OBJECTS TRACKING

Tracking mobile objects in distributed environments is an important task in a number of areas like air and road traffic, infrastructure protection, national and international crime, or missile defense. The example here relates to tracking aerial objects by a dynamic network of unmanned aerial vehicles, UAVs (Sapaty, 2008), with the following features to be taken into account.

- Each UAV can observe only a limited part of space.
- To keep the whole observation continuous, the object discovered should be handed over between neighboring UAVs during its movement, along with the data accumulated about it.
- The model can catch each object and accompany it individually by the mobile intelligence, while propagating between the WPL interpreters in UAVs.
- Many such objects can be picked up and chased in parallel by a dynamic UAV network.

The following program, starting in all units, catches the object it sees and follows it wherever it goes, if it is not seen from this point any more (its visibility becomes lower than a given threshold).

```
hop all_nodes; Frontal Threshold = 0.1;
frontal Object =
  select_max_visible (aerial, Threshold);
repeat (
  loop (visibility (Object) > Threshold );
  choose_destination_with_max_value (
    hop all_neighbors.
    visibility (Object) > Threshold))
```

A snapshot of a possible situation in a distribute space is shown in Fig. 22. The information about the tracked objects can be accumulated by individual mobile intelligences (Sapaty, Corbin, Seidensticker, 1995), which can cooperate with each other, making individual or collective decisions about the further fate of the objects (e.g. classifying them as friendly or hostile).

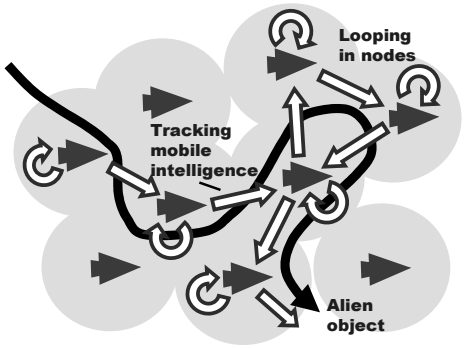



Figure 22: Collective tracking of a mobile object.

# 12 COLLECTIVE BEHAVIOR

The higher-level, semantic WPL scenarios are well understandable by humans, who can perform jobs written in the language and delegate other jobs to other group members, establishing runtime relations with each other. These scenarios also represent fully formal descriptions that can be effectively interpreted by robots and their groups automatically.

Both human and robotic suitability allow for a fully unified approach to organization of teams that can range from purely human to purely robotic. These teams can be open and emergent, and can operate in unpredictable environments, where team members can indiscriminately fail at any time but the mission scenario, collectively interpreted by the distributed group, can survive and fulfill objectives. The collective team behavior can be based on a loose organization like swarms, or can be strictly and hierarchically controlled. Different solutions in WPL throughout this organizational range are possible, including any combined ones (Sapaty, 2005).

With the initial distribution of units shown in Fig. 23, let us consider a collective swarm-like movement, where each unit randomly, within certain hop limits defining general direction, tries to move in new positions, keeping the established threshold distance to other units.

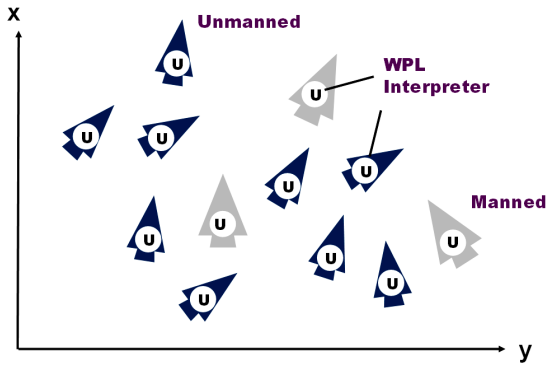



Figure 23: Initial distribution of units.

This can be done by the following program, which can start from any unit, manned or unmanned.

```
nodal (Limits, Range, Shift);
hop all_nodes;
Limits = (dx (0, 8), dy (- 2, 5)); Range = 5;
repeat (
  Shift = random (Limits );
  if empty hop (Shift, Range) then move Shift)
```

A snapshot of the group movement by this spatial program is depicted in Fig.24.

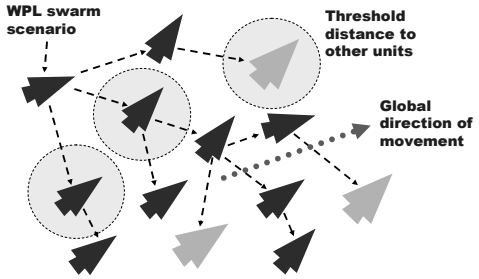



Figure 24: A swarm movement snapshot.

To have more coordinated actions of the group, we may set up a distributed hierarchical infrastructure over it, to be used in command and control and in maintaining global awareness. As the group is distributed in space and distances between units can change, such an infrastructure should be preferably based on the current physical position of the units, with top of the hierarchy to be close to the group's center, in order to optimize global coordination. We will consider here how the topologically central unit can be found at runtime, during the movement within a swarm, and how the C2 hierarchy can be formed starting from this central unit. The following distributed program, starting from any unit, finds topologically central unit of the distributed swarm, which is shown in Fig. 25.

```
frontal Aver =
  average (hop all_nodes;  WHERE);
nodal Center =
  element (
    min (
      hop all_nodes;
      distance (Aver, WHERE) & ADDRESS), 2)
```

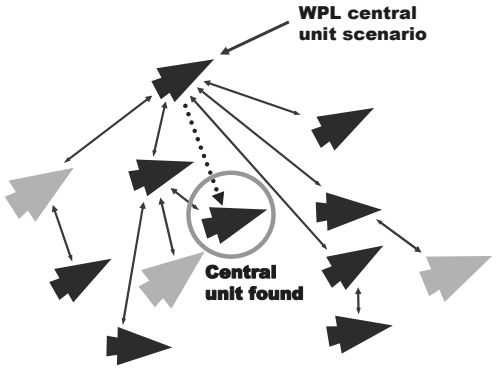



Figure 25: Finding topologically central unit.

Starting from the central unit found, the next program creates runtime hierarchical infrastructure with oriented links infra, as shown in Fig. 26.

```
frontal Range = 20.
  repeat (
    create_links (
      + infra, first_come, nodes (Range)))
```

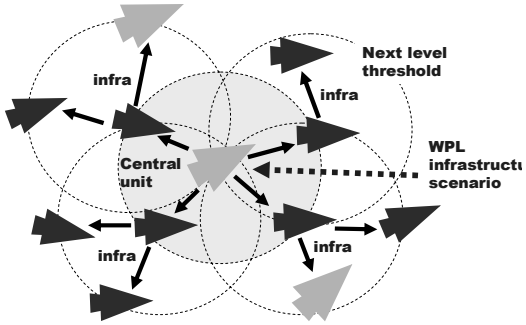



Figure 26: Hierarchical infrastructure built.

This runtime hierarchy created may be effectively used for maintaining global awareness in the distributed space, collection and fusion of targets seen by individual units, spreading the set of collected targets back to all units, which may select the most

suitable ones for an individual impact. The following program, navigating the infrastructure created, follows this scenario, as shown in Fig. 27.

```
repeat (
  if nonempty (
    frontal Seen = Repeat (
              Free (detect targets),
              Hop_links + infra)) then
    repeat (
      free (if TYPE == UAV then
               select_move_shoot Seen),
      hop_links + infra))
```

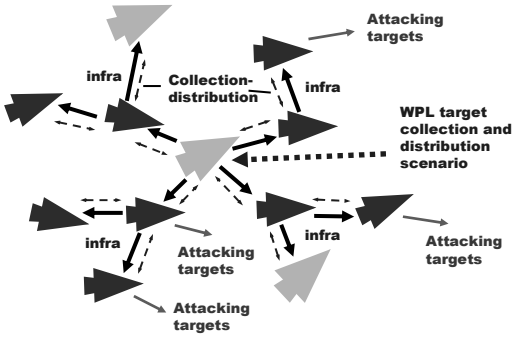



Figure 27: Hierarchical fusion and distribution of targets.

As the whole group moves and distances between separate units may change in the swarm, the programs of finding the center and hierarchical infrastructure may be repeated with a certain regularity, which will help to maintain the group's optimal spatial organization in a distributed environment. Any position in this dynamic hierarchy (the top one including) may happen to be occupied by any unit at any moment of time, regardless of whether it is manned or unmanned.

Many more applications of this world processing paradigm (previously known as WAVE) can be found in (Sapaty, 1999, 2005), also (Sapaty, Morozov, Finkelstein, et al., 2007).

# 13 CONCLUSIONS

We have touched only some of the areas currently in active investigation for the WP technology being developed. The experience obtained allows us to claim the following.

- The proposed technology converts any distributed system into a universal spatial computer capable of solving complex problems on itself and on the surrounding environment.
- This system, for example, can be a single unit or a group (or army) of them, with individual units being manned or unmanned.
- The whole system is driven by high-level scenarios setting how to behave as a whole and what to do, while omitting traditional implementation details which are effectively delegated to intelligent distributed interpretation system.
- The system scenarios in the World Processing Language are very compact and can be created at runtime, on the fly, swiftly reacting on a rapidly changing environment and mission goals.
- Any scenario can start from any available component and cover the system at runtime, during its evolution.
- The approach may offer real possibilities for a runtime recovery after indiscriminate damages, including reassembling of the whole system (or what remains of it) from any point.
- The technology can help dominate over other distributed system organizations, especially those explicitly based on communicating and interacting parts (agents).

# REFERENCES


Sapaty, P. S., 1993. *A distributed processing system*, European Patent No. 0389655, Publ. 10.11.93, European Patent Office.

Sapaty, P. S., Corbin, M. J., Seidensticker, S., 1995. Mobile intelligence in distributed simulations, *Proc. 14th Workshop on Standards for the Interoperability of Distributed Simulations*, IST UCF, Orlando, FL, March.

Sapaty, P. S., 1999. *Mobile Processing in Distributed and Open Environments*, John Wiley & Sons, ISBN: 0471195723, New York, February, 436p.

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, 9p.

Sapaty, P. S., 2005. *Ruling Distributed Dynamic Worlds*, John Wiley & Sons, New York, May, 256p, ISBN 0-471-65575-9

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., Morozov, A., Finkelstein, R., Sugisaka, M., Lambert, D., 2007. A new concept of flexible organization for distributed robotized systems. *Proc. Twelfth International Symposium on Artificial Life and Robotics (AROB 12th'07)*, Beppu, Japan, Jan 25-27, 8p.

Sapaty, P., Morozov, A., Sugisaka, M., 2007. DEW in a network enabled environment, *Proc. international conference Directed Energy Weapons 2007*, Feb. 28 - March 1, Le Meridien Piccadilly, London, UK.

Sapaty, P., 2007. Global management of distributed EW-related systems, *Proc. Electronic Warfare: Operations & Systems 2007*, 19-20 Sept., Thistle Selfridge, London, UK.

Sapaty, P., 2007a. 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., 2008. Grasping the whole by spatial intelligence: A higher level for distributed avionics, *Proc. international conference Military Avionics 2008*, Jan. 30 - Feb.1, Café Royal, London, UK.

## BRIEF BIOGRAPHY

Dr. Peter Simon Sapaty, educated as power networks engineer, is with distributed systems for 40 years, implementing heterogeneous computer networks from the end of the sixties. Being chief research scientist and director of distributed simulation and control at the Institute of Mathematical Machines and Systems, National Academy of Sciences of Ukraine, also worked in Czechoslovakia, Germany, UK, Canada, and Japan as project leader, research professor, department head, and special invited professor; chaired a special interest group on mobile cooperative technologies within Distributed Interactive Simulation project in the US. Peter invented and prototyped a distributed networking technology (supported by Siemens/Nixdorf, Ericsson UK, and Japan Society for the Promotion of Science) used in different countries and resulted in a European Patent and two John Wiley books. His interests include models and languages for coordination and simulation of distributed dynamic systems with application in intelligent network control, emergency management, infrastructure protection, and cooperative robotics.

# BEHAVIORAL DEVELOPMENT FOR A HUMANOID ROBOT

## *Towards Life-Long Human-Robot Partnerships*

Ronald C. Arkin
*Mobile Robot Laboratory, College of Computing, Georgia Tech, Atlanta, GA, U.S.A. 30332*
*arkin@cc.gatech.edu*

## EXTENDED ABSTRACT

A significant research effort was conducted at Sony's Intelligence Dynamics Laboratory (SIDL), involving personnel from Georgia Tech, MIT, CMU, Osaka University, and SIDL, working towards the implementation of a theory of designed development for a humanoid robot. This research involves numerous insights gleaned from cognitive psychology (drawn from both new and old theories of behavior) and integrating these techniques into Sony's humanoid robot QRIO architecture with the long-term goal of providing highly satisfying longterm interaction and attachment formation by a human partner. Included are models of deliberative (willed) reasoning and its interfacing with a reactive (automatic) controller (Glasspool 00, Shallice and Burgess 96, Ulam and Arkin 07). In particular aspects of skill transference from planned to routine activity are incorporated (Cooper and Glasspool 01, Cooper and Shallice 97, Chernova and Arkin 07). In addition, a multi-method learning technique inspired by assimilation models of Piaget provides for runtime incorporation of disparate learned skills into the existing behavioral substrate (Takamuku and Arkin 07). Finally non-verbal communication mechanisms that overlay ongoing behavior performance and utilize both proxemics (spatial separation) and kinesics (body language) are described (Brooks and Arkin 07). All of the underlying models, their implementation and the results obtained on QRIO are presented.

## REFERENCES

Brooks, A. and Arkin, R.C., "Behavioral Overlays for Non-Verbal Communication Expression on a Humanoid Robot", *Autonomous Robots*, Vol. 22, No.1, pp. 55-75, Jan. 2007.

Chernova, S. and Arkin, R.C., "From Deliberative to Routine Behaviors: A Cognitively-Inspired Action Selection Mechanism for Routine Behavior Capture", *Adaptive Behavior*, Vol. 15, No. 2, pp. 199-216, June 2007.

Cooper, R., & Glasspool, D., "Learning action affordances and action schemas", *Connectionist Models of Learning, Development, and Evolution*, 133-142, 2001.

Cooper, R., and Shallice, T., "Modeling the selection of routine action: Exploring the criticality of parameter values", in *Proceedings of the 19th annual conference of the cognitive science society* p. 130-135, 1997.

Glasspool, D., "The integration of control and behavior: Insights from neuroscience and AI", in *Proceedings of the How to Design a Functioning Mind Symposium at AISB-2000*, pp. 77–84, 2000.

Shallice, T. and Burgess, P., "The domain of supervisory processes and temporal organization of behavior", in *Philosophical Transactions of the Royal Society of London B,* vol. 351, pp. 1405–1412, 1996.

Takamuku, S. and Arkin, R.C., "Multi-method Learning and Assimilation", *Robotics and Autonomous Systems*, Vol. 55, No. 8, pp. 618-627, 2007.

Ulam, P. and Arkin, R.C., "Biasing Behavioral Activation with Intent", to appear in *Intelligent Service Robotics*, 2008.

## BRIEF BIOGRAPHY

Ronald C. Arkin is Regents' Professor and the Director of the Mobile Robot Laboratory in the College of Computing at the Georgia Institute of Technology. He has held visiting positions at the Royal Institute of Technology in Stockholm, the Sony Intelligence Dynamics Laboratory in Tokyo, and LAAS/CNRS in Toulouse. Dr. Arkin's research interests include behavior-based reactive control and action-oriented perception for mobile robots and unmanned aerial vehicles, hybrid deliberative/ reactive software architectures, robot survivability, multiagent robotic systems, biorobotics, human-robot interaction, robot ethics, and learning in autonomous systems. He has over 130 technical publications in these areas and has written a textbook entitled Behavior-Based Robotics and is the Series Editor for the MIT Press book series

Intelligent Robotics and Autonomous Agents. Prof. Arkin served two terms on the Administrative Committee of the IEEE Robotics and Automation Society, serves as the co-chair of the IEEE RAS Technical Committee on Robot Ethics, and also served on the National Science Foundation's Robotics Council. He was elected a Fellow of the IEEE in 2003, and is a member of AAAI and ACM.

# SWARM INTELLIGENCE AND SWARM ROBOTICS
## *The Swarm-Bot Experiment*

Marco Dorigo
*IRIDIA, Université Libre de Bruxelles*
*Belgium*

Abstract: Swarm intelligence is the discipline that deals with natural and artificial systems composed of many individuals that coordinate using decentralized control and self-organization. In particular, it focuses on the collective behaviors that result from the local interactions of the individuals with each other and with their environment. The characterizing property of a swarm intelligence system is its ability to act in a coordinated way without the presence of a coordinator or of an external controller. Swarm robotics could be defined as the application of swarm intelligence principles to the control of groups of robots. In this talk I will discuss results of Swarm-bots, an experiment in swarm robotics. A swarm-bot is an artifact composed of a swarm of assembled s-bots. The s-bots are mobile robots capable of connecting to, and disconnecting from, other s-bots. In the swarm-bot form, the s-bots are attached to each other and, when needed, become a single robotic system that can move and change its shape. S-bots have relatively simple sensors and motors and limited computational capabilities. A swarm-bot can solve problems that cannot be solved by s-bots alone. In the talk, I will shortly describe the s-bots hardware and the methodology we followed to develop algorithms for their control. Then I will focus on the capabilities of the swarm-bot robotic system by showing video recordings of some of the many experiments we performed to study coordinated movement, path formation, self-assembly, collective transport, shape formation, and other collective behaviors..

## BRIEF BIOGRAPHY

Marco Dorigo received the Laurea (Master of Technology) degree in industrial technologies engineering in 1986 and the doctoral degree in information and systems electronic engineering in 1992 from Politecnico di Milano, Milan, Italy, and the title of Agrégé de l'Enseignement Supérieur, from the Université Libre de Bruxelles, Belgium, in 1995. From 1992 to 1993 he was a research fellow at the International Computer Science Institute of Berkeley, CA. In 1993 he was a NATO-CNR fellow, and from 1994 to 1996 a Marie Curie fellow. Since 1996 he has been a tenured researcher of the FNRS, the Belgian National Fund for Scientific Research, and a research director of IRIDIA-CoDE, the artificial intelligence laboratory of the Université Libre de Bruxelles. He is the inventor of the ant colony optimization metaheuristic. His current research interests include swarm intelligence, swarm robotics, and metaheuristics for discrete optimization. Dr. Dorigo is the Editor-in-Chief of the Swarm Intelligence journal. He is an Associate Editor for the IEEE Transactions on Evolutionary Computation, the IEEE Transactions on Systems, Man, and Cybernetics, and the ACM Transactions on Autonomous and Adaptive Systems. He is a member of the Editorial Board of numerous international journals, including: Adaptive Behavior, AI Communications, Artificial Life, Cognitive Systems Research, Evolutionary Computation, Information Sciences, Journal of Heuristics and Journal of Genetic Programming and Evolvable Machines. In 1996 he was awarded the Italian Prize for Artificial Intelligence, in 2003 the Marie Curie Excellence Award, and in 2005 the Dr A. De Leeuw-Damry-Bourlart award in applied sciences. He is a fellow of the IEEE and of the ECCAI, the European Coordinating Committee for Artificial Intelligence.

# SIGNAL PROCESSING, SYSTEMS MODELING AND CONTROL

# FULL PAPERS

# DESIGN OF AN ANALOG-DIGITAL PI CONTROLLER WITH GAIN SCHEDULING FOR LASER TRACKER SYSTEMS

Christian Wachten, Lars Friedrich, Claas Müller, Holger Reinecke
*Department of Microsystems Technology, University of Freiburg, Georges-Koehler-Allee 103, 79110 Freiburg, Germany*
*wachten@imtek.de, lfriedri@imtek.de, clmuelle@imtek.de, reinecke@imtek.de*

Christoph Ament
*Institue for Automation and Systems Engineering, TU Ilmenau, 98693 Ilmenau, Germany*
*christoph.ament@tu-ilmenau.de*

Keywords: Laser tracker system, *PI* controller with μC, Analog-digital design, Absolute distance measurement.

Abstract: Laser trackers are important devices in position metrology. A moving reflector is tracked by a laser beam to determine its position in space. To ensure a proper function of the device the feedback control loop is an essential part. An analog *PI* controller with online parameter adaptation and absolute distance measurement ability is used to guarantee an optimal dynamic system. The feedback controller is connected to a quadrant detector which serves as the sensor element in the control loop. The position of an incoming laser beam is measured by the quadrant detector and the controller provides the input signals for a subsequent actuator. The control variable is the deviation of the laser beam from the centre of the diode which should ideally be zero. The actuator consists of two axes and each one is equipped with a rotatable mirror. The task of the controller is to rotate the mirrors in such a way so that the laser beam follows the movements of the reflector. To design an optimal controller linear, time-invariant models of the actuator and the position sensor are developed to optimize its parameters. The gain of the plant correlates with the distance between the reflector and the laser tracker. To achieve the optimal dynamic performance the controller is automatically adapted to the distance during operation. A method based on oscillation injection to measure the absolute distance is developed. Due to higher dynamic demands a standard analog *PI* controller is implemented with the controller gain tuned by digital potentiometers. A microcontroller is used to adjust the parameters and to estimate the distance. During the power up sequence and in case of a beam loss the system is completely controlled by the digital part.

## 1 INTRODUCTION

Laser trackers are devices which are used in position metrology and in calibration tasks due to their capability of doing static as well as dynamic high accuracy measurements (Riemensperger & Gottwald 1990). A HeNe laser with a Gaussian beam profile emits two light beams with different frequencies $f_1$ and $f_2$. These beams are divided by an interferometer into a reference beam and a measurement beam. The measurement beam leaves the interferometer and is deflected by the mirrors of an actuator in such a manner that it follows the movement of a retroreflector. The reflected light is analyzed by a position sensitive detector, the analog output signals of which are used to determine the position of the incoming beam. The reflected light beam also interferes with the reference beam in the interferometer. So, by measuring the two mirror angles of the actuator and the relative distance given by the interferometer the position of the reflector is calculated by using an analytical model. Figure 1 shows the operation principle.

An important part of the tracker is the feedback controller in the tracking unit. It provides the input signals for the actuator, so that the laser beam follows the movement of the retroreflector. The basic task of the controller is to guarantee the proper interferometer function. Dynamic aspects like a high velocity or high acceleration of the retroreflector with a low contouring error are also important.

We present the development of a fast and cost effective analog feedback controller for a tracking unit that can be used with laser tracker systems. The

tracking unit is designed for working with an interferometer but can also act as an autonomous system because of the integrated absolute distance measurement technique. A distance of about eight meters between the tracker unit and the reflector is easily achieved in experiment without static tracking errors. Furthermore, the lateral offset of the laser beam does not exceed a quarter of the beam diameter and thus guarantees stable interferometer functionality.

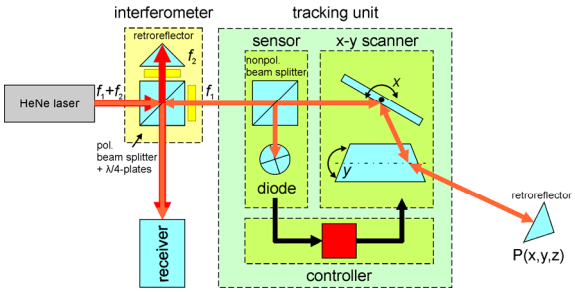



Figure 1: Operation principle of a laser tracker system. The frequency $f_1$ represents the measurement beam and $f_2$ the reference beam.

## 2 SYSTEM COMPONENTS

The tracking unit consists of three components (see figure 1). The first component is the sensor element which is a combination of a nonpolarizing beam splitter and a quadrant diode. The second component is a x-y scanner with two magnetically driven mirrors that deflect the laser beam (galvanometer scanner). The third component is the controller that connects the sensor element with the actuator.

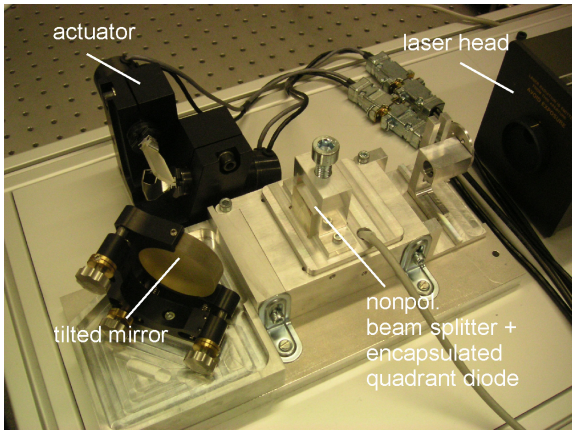



Figure 2: Photograph of the tracking unit. The two mirrors of the actuator deflect the incoming laser beam. A tilted mirror is used to adjust the light path.

The laser beam hits a nonpolarizing beam splitter. It has a division ratio of 50:50. Afterwards, it is deflected by the scanner and is then reflected by the retroreflector. The retroreflector has the unique property that the incoming laser beam is reflected into the same direction where it came from. On its returning path the reflected beam hits the nonpolarizing beam splitter again and a part of the beam is deflected on a quadrant diode. This diode provides the analog input voltages for the feedback controller since it has an integrated transimpedance amplifier. The feedback controller generates the input signals for the scanner. Ideally, the laser beam is centered on the diode and the position signals equal zero volts. By moving the reflector the laser beam leaves the center on the diode and the position signals change. The controller compensates the position change and modifies the input signals of the actuator. The two mirrors rotate and deflect the laser beam in such a way that the offset becomes zero. Figure 2 shows a photograph of the complete tracking unit.

### 2.1 Laser Head

The laser is a class II HeNe laser with a Gaussian beam profile. It emits two linear, orthogonal polarized beams with a split frequency of about 1.8 MHz. The beam diameter is about 6 mm. The power $P$ of the laser is about 120 µW.

### 2.2 Magnetic Actuator

The magnetic actuator is a galvanometer scanner produced by Cambridge Technology. It has silver coated mirrors which allow a maximal beam aperture of 10 mm. Each mirror is magnetically driven and has its own analog *PID* controller to hold the desired position. The transfer factor is 0.83 V/° (mechanical) at the input side of the controller and 0.5 V/° (mechanical) at the output side of the position detector. Integrated sensors allow the measurement of the rotation angle of the mirrors. The short term stability is about 8 µrad. The maximal mechanical rotation angle is ± 12.5° limited by the used assembly.

### 2.3 Quadrant Photodetector

The sensor element of the tracking unit is a quadrant photodiode. Figure 3 shows a photograph of the photodiode and a sketch of its quadrants. Each quadrant is sensitive to light with a sensitivity that is specified to 0.54 A/W at a wavelength of 900 nm. The spacing

between the quadrants is 0.2 mm, the quadrant radius is 3.99 mm.

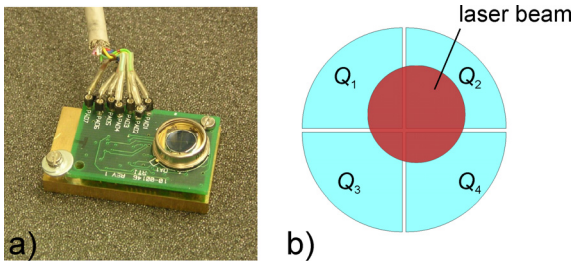



Figure 3: Photograph of the quadrant detector a) and sketch b) of its quadrants. The voltage level of the position signals depends on the area that is covered by the laser beam, its shape, its power and the wavelength.

The diode current depends on the power of the laser beam, its shape, its wavelength and the area that is covered. To perform a current to voltage transformation a transimpedance amplifier is put on the same circuit. The output of the amplifier consists of three voltages that completely define the position and the power of the laser beam. The signals can be calculated with the following formulas:

$$V_{Sum} = (I_1 + I_2 + I_3 + I_4) \cdot 10^4 \text{ V/A} \quad (1)$$

$$V_{TB} = ((I_1 + I_2) - (I_3 + I_4)) \cdot 10^4 \text{ V/A} \quad (2)$$

$$V_{LR} = ((I_1 + I_3) - (I_2 + I_4)) \cdot 10^4 \text{ V/A} \quad (3)$$

The symbol $I_i$ represents the current of a quadrant $Q_i$. The signal $V_{Sum}$ is the summation of the voltages that are generated by each quadrant and thus is an indicator for the total power of the incoming light beam at a known wavelength. The signal $V_{TB}$, the so called top-bottom voltage, represents the position of the laser beam in vertical direction. The signal $V_{LR}$, the so called left-right voltage, represents the position of the laser beam in horizontal direction. If for example the laser power is the same on each quadrant, $V_{TB}$ and $V_{LR}$ become zero volts.

# 3 MODELING OF THE PLANT

A block diagram of the system is shown in fig-ure 4. The output signal $y$ of the system is the deviation of the light beam from the center of the quadrant diode. This deviation should become zero for each component. The output signal $y$ is a superposition of the movement of the reflector and the compensation part of the actuator. The gain between the mirror angle and the movement of the light beam on the diode depends on the distance between the reflector and the tracking unit. This is modeled in the block "light path". The symbol $z$ represents the position of the reflector in space and thus is a three-dimensional vector. The signal $y$ represents the beam position on the diode area and thus is a two-dimensional vector.

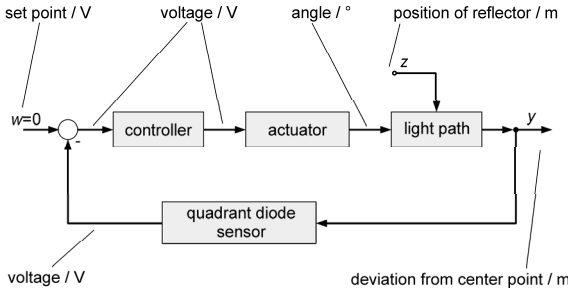



Figure 4: Block diagram of the plant. The different inputs and outputs with their units are shown.

To design an optimal feedback controller, models are developed and the model parameters are identified for the blocks "actuator", "quadrant diode" and "light path".

## 3.1 Modeling of the Block "actuator"

To obtain a transfer function for the magnetic actuator the step response is recorded for each axis monitoring the position output of the integrated angle encoders. A square wave with a peak-peak voltage of 100 mV (corresponding to an angle of about 0.06°) and a frequency of 30 Hz is applied to the inputs of the scanner.

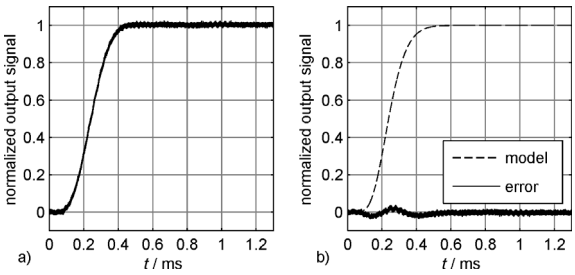



Figure 5: Normalized and averaged measurement a) of the step response of the y mirror and the model b) of the step response with its error. The step rises at t = 0 s.

Figure 5 shows the measured step response and the step response of the model for the y mirror. It is assumed that the scanner has $PT_n$ behavior and thus can be modeled with a $PT_n$ element in (4).

$$H_i(s) = \frac{K_i}{(1 + T_i \cdot s)^{n_i}}, \; i = x,y \quad (4)$$

The parameter $K_i$ describes the gain and the parameter $T_i$ represents the time constant for axis $i$. The parameter $n_i$ stands for the order of the element. To obtain the parameters the method of the time percentage values is applied. Since the measurement is not supposed to have ideal $PT_n$ behavior a least-squares fit is done to optimize the parameters so that the error between the model and the measurement becomes minimal. The start parameters for the optimization are the results given by the method of the time percentage values (Schwarze 1962). The final optimization yields in $n_y = 9$ and $T_y = 27.51$ µs for the y mirror. The parameters for the x mirror result in $n_x = 10$ and $T_x = 23.42$ µs. Taking into account that the gain between the output signal and the mechanical deflection is 0.5 V/° the gains are calculated to $K_x = 1.210$ °/V and $K_y = 1.207$ °/V, respectively. The -3 dB frequency is about 1.8 kHz (model).

## 3.2 Modeling of the Block "diode"

The time response of the diode can be modeled in the same way as the time response for the actuator. A red LED is used to generate the step response because it is fast enough and its time response can be neglected. The output signal $V_{Sum}$ is measured. The LED has a power of about 7 µW. Figure 6 shows the measurement and the model of the step response.

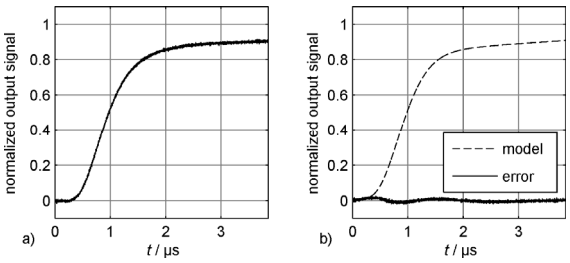



Figure 6: Normalized and averaged measurement a) of the step response of the diode and the model b) with its error. A red, pulsed LED is used to illuminate the active area. The step rises at t = 0 s.

The modeling with only a single $PT_n$ element is not applicable because there is a steep increase of the signal until 1.6 µs. Afterwards, the signal increases very slowly and does not reach 100% even after 4 µs. Therefore, it can be shown that a good approximation is a combination of a $PT_n$ element and a $PT_1$ element. This is done in (5).

$$H_D(s) = \frac{g}{(1+T_1 \cdot s)^{n_D}} + \frac{1-g}{1+T_2 \cdot s} \qquad (5)$$

The parameters $T_1$ and $T_2$ represent time constants of the two elements, the parameter $g$ normalizes the output and the parameter $n_D$ represents the order of the $PT_n$ element. All parameters are obtained using a least squares fit so that the deviation of the model and the measurement becomes minimal. The optimal parameters are $n_D = 7$, $T_1 = 136$ ns, $T_2 = 4.75$ µs, $g = 0.793$. The model predicts a -3 dB frequency of 250 kHz.

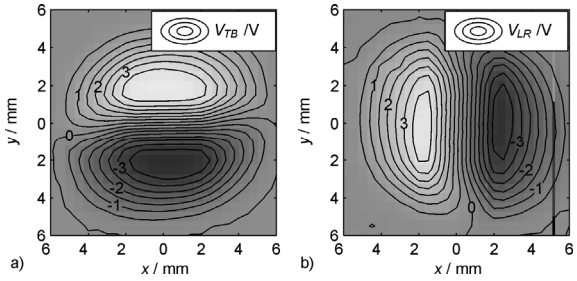



Figure 7: Local behavior of the diode. There is a nonlinear relation between the output voltage and the position of the beam.

Because the diode generates position signals not only the time response is important but also its local behavior. Figure 7 shows the local behavior of diode. The quadrant diode was put onto an x-y station and a laser diode with a power of 771 µW was installed in front of it. The station moves to 900 defined positions that are placed in an equally spaced square.

In figure 7 it can be seen that there is a nonlinear relation between the position and the output voltages. In the center of the diode the contour lines are nearly parallel to the corresponding axis and so a linearization is possible. It is obvious that the signal $V_{TB}$ only depends on a movement in y direction and $V_{LR}$ only depends on a movement in x direction. As a result, each axis of the scanner can be regarded as independent.

## 3.3 Modeling of the Block "light path"

The block "light path" depends on the position of the reflector. If the angular errors are neglected the reflector can be approximated as its center point in a plane (see figure 8a). The hitting point of the incoming laser beam is point reflected with the center. An offset between the incoming and the reflected laser beam can have different reasons, for example a rotation or a lateral displacement of the reflector.

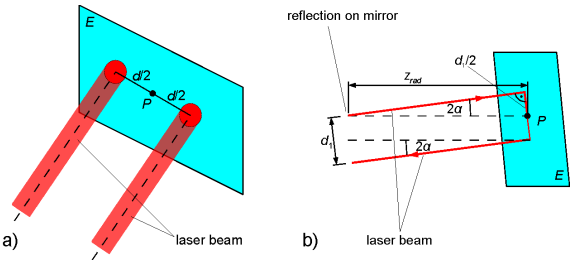



Figure 8: Modeling of the reflector. An incoming beam is reflected at the center of the reflector.

The first effect is shown in figure 8b. The mirror rotation angle is $\alpha$. It is assumed that the reflector is $z_{rad}$ away from the tracking unit. So the offset $d_1$ can be written as

$$d_1 = 2 \cdot z_{rad} \cdot \sin(2\alpha) \approx 4 \cdot z_{rad} \cdot \alpha \qquad (6)$$

The second effect resulting in a beam offset is the lateral displacement $z_{lat}$ of the reflector. The indices represent the coordinate frame of the diode. So it can be written as

$$d_{2,i} = 2 \cdot z_{lat,i}\,,\ i = x,y \qquad (7)$$

The total deflection on the quadrant diode is a combination of these two effects as shown in (8)

$$y_i = d_1 + d_{2,i}\,,\ i = x,y. \qquad (8)$$

# 4 CONTROLLER DESIGN

The controller has to be designed separately for the x and the y-axis. Because of the decoupling of the axes shown in figure 7 the problem is reduced to the controller design for one axis. Exemplarily, the x-axis is used to demonstrate the design process. It can be derived from the block "light path" that the gain of the plant correlates with the distance $z_{rad}$. In a first step it is assumed that $z_{rad}$ is constant. During operation the controller should be automatically adapted to the distance so the restriction $z_{rad}$ = const. is dropped. This adaptation to the distance is known as gain scheduling.

The controller has to fulfill several requirements for the use in laser tracker systems. First, the offset of the beam should be smaller then a quarter of the beam diameter to guarantee the interferometer function. Second, a high velocity of the reflector is necessary to allow rapid movements of the object. The third requirement is the robustness against vibrations without any disturbance of the measurement accuracy. Of course, a large measurement volume is desirable, too.

To achieve stationary accuracy the controller should possess an integrating part because the plant only consists of proportional blocks. A pure *I* controller is also possible but a proportional part enhances the performance (Merz & Jaschek 1996). A *PI* controller is well suited for plants with $PT_n$ behavior.

A standard *PI* controller is proposed due to its simple design and to reduce the analog circuit complexity because the derivative part is missing. The transfer function is well known and given in (9).

$$H_R(s) = K_R \cdot \left(1 + \frac{1}{T_R \cdot s}\right) \qquad (9)$$

The parameter $K_R$ describes the gain of the controller and $T_R$ its time constant. First, these parameters are determined by the classic frequency response method as described by Föllinger (1994). The time constant $T_R$ is set to $T_x$ = 23.42 µs because this is the dominating time constant of the x mirror in the plant. The gain $K_{R,30°}$ is set to 0.1231 to obtain a phase margin of 30° for the gain crossover frequency. It is of interest to examine the disturbance transfer function because the set point ($w(t)$ = 0) is constant. $H_z(s)$ is calculated in (10) and is derived from the block diagram in figure 4.

$$H_z(s) = \frac{2}{1 + H_S(s) \cdot H_R(s) \cdot H_D(s)} \qquad (10)$$

To estimate the performance of the classic feedback controller the response to a ramp in $z_{lat,i}$ and the magnification factor is simulated for the disturbance transfer function in (10).

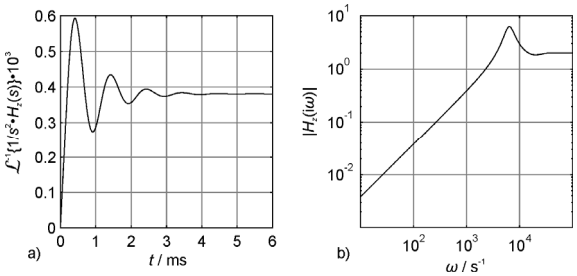



Figure 9: Simulation of the ramp response a) and the frequency response b) of the disturbance transfer function with a classic *PI* controller.

A ramp is chosen because it is the strongest requirement concerning the reflector movement. A

step is not applicable because in reality the position of the retroreflector cannot rapidly change.

Figure 9a) shows the response on a ramp in $z_{lat}(t)$ at the output $y_x$ of the system. The maximal value of the overshoot at the output $y_x$ is $0.60 \cdot 10^{-3}$ m until a constant contouring error of $0.38 \cdot 10^{-3}$ m is reached. Figure 9b) shows the frequency response. Low frequencies are damped due to the integrating part. But there is a magnification factor of 6.3 at a frequency of 1 kHz.

The maximal value is reached at the overshoot, but the contouring error is much lower. Therefore, the maximum of the ramp response has to be minimized so that the overshoot is reduced at the cost of the contouring error. The aim is the adaptation of the overshoot to the contouring error. For $\omega \to \infty$, $|H_z(i\omega)|$ converges to 2, so an arbitrary factor of 4 is proposed for all frequencies to guarantee robustness against vibrations. To identify the optimal controller parameters $K_R$ and $T_R$ were varied in a range of $K_R = 0.1 \cdot K_{R,30°} \ldots 10 \cdot K_{R,30°}$ and $T_R = 0.1 \cdot T_x \ldots 10 \cdot T_x$. The raster was $\Delta K_R = 0.05 \cdot K_{R,30°}$ and $\Delta T_R = 0.05 \cdot T_x$. Figure 10 shows the ramp response and frequency response for the optimized parameters $K_{R,opt} = 0.542$ and $T_{R,opt} = 135$ µs. The maximal value of the response is about $0.51 \cdot 10^{-3}$ m with a very low overshoot and the magnification factor does not exceed 4. So, the quality of control is much better in comparison with the classic approach.

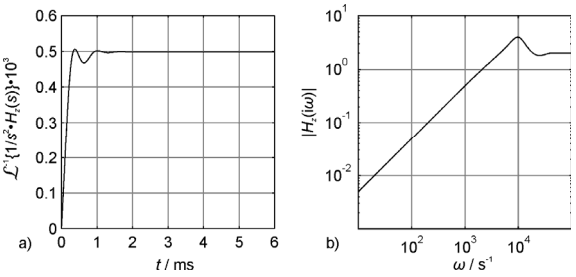



Figure 10: Simulation of the ramp response a) and the frequency response b) of the disturbance transfer function with optimal controller parameters in regard to contouring error and magnification factor.

It was assumed that the total gain is focused in the parameter $K_R$. This is not the case in a real system. The real total gain is a product of the controller gain, the mirror gain, the light path and the sensitivity of the quadrant diode. It can be written as

$$K_{R,opt} = K_R \cdot K_x \cdot 4 \cdot z_{rad} \cdot K_D \tag{11}$$

$$\Leftrightarrow K_R = \frac{K_{R,opt}}{K_x \cdot 4 \cdot z_{rad} \cdot K_D} \tag{12}$$

To adapt the controller gain, the parameters $K_R$ and $K_D$ have to be updated during operation (gain scheduling). The parameter $K_D$ is obtained by measuring the voltage $V_{Sum}$ of the quadrant diode. The parameter $z_{rad}$ has to be estimated.

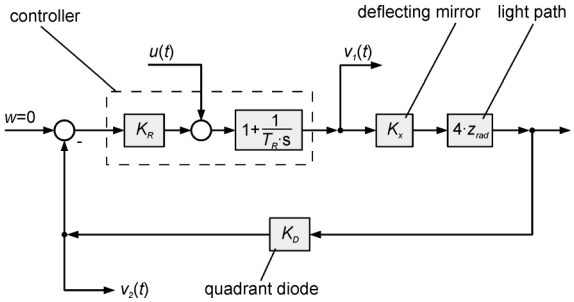



Figure 11: Control loop with introduced signal $u(t)$ to estimate $z_{rad}$. The signals $v_1(t)$ and $v_2(t)$ can be measured and depend on $z_{rad}$.

To estimate $z_{rad}$ a known signal $u(t)$ is introduced into the control loop (figure 11). Only the spectral components are considered which are below the cut-off frequency of the diode. So, it is possible to simplify the transfer functions and consider only their proportional parts. But the time response of the controller cannot be neglected because there is a strong dependency on low frequencies introduced by the integrating part. The multiplication with $z_{rad}$ is modeled as proportional part because $z_{rad}$ changes slowly in comparison to the system dynamic.

The signals $v_1(t)$ and $v_2(t)$ depend on the distance $z_{rad}$. The transfer functions can be calculated with (13).

$$H_i(s) = \frac{\mathcal{L}\{v_i(t)\}}{\mathcal{L}\{u(t)\}}, \; i = 1,2 \tag{13}$$

It can be shown that there is a higher sensitivity of $|H_1(i\omega_0)|$ to a change in $z_{rad}$ if signal $v_1(t)$ is measured. With (13) and figure 11 the transfer function $H_1(s)$ is calculated in (14).

$$H_1(s) = \frac{1 + \frac{1}{T_R \cdot s}}{1 + \left(1 + \frac{1}{T_R \cdot s}\right) \cdot K_R \cdot K_D \cdot 4 \cdot z_{rad} \cdot K_x} \tag{14}$$

The signal $v_1(t)$ is a superposition of the movement of the reflector and the introduced signal $u(t)$. A si-

nusoidal signal with a frequency $\omega_0$ is proposed. So, a subsequent band-pass filter with the center frequency $\omega_0$ suppresses the disturbance signal introduced by the reflector. To estimate the distance $z_{rad}$ the amplitude of $u(t)$ is compared with the amplitude $v_1(t)$ and so $|H_1(i\omega)|$ is calculated. With the experimentally obtained relationship for $K_D$ in (15), $z_{rad}$ is calculated in (16). The proportional value $m$ has a value of 45.13 $V^{-1}m^{-1}$.

$$K_D = \frac{m \cdot V_{Sum}}{4 \cdot K_x} \tag{15}$$

$$z_{rad} = \left(\sqrt{A^2(1-(A^2-2)T_R^2\omega_0^2 + T_R^4\omega_0^4)} - A^2T_R^2\omega_0^2\right)/\left(A^2K_R m V_{Sum}(1+T_R^2\omega_0^2)\right) \tag{16}$$

with $A = |H_1(i\omega_0)|$.

To obtain an appropriate value for the amplitude of the signal $u(t)$ the beam deviation $y_x(t)$ from the center of the quadrant diode is analyzed. The deviation should be smaller than a quarter of the beam diameter to guarantee the interferometer function. During distance estimation the deviation $y_x(t)$ is a superposition of the lateral movement of the reflector $z_{lat,x}(t)$ and the introduced signal $u(t)$. Therefore, the amplitude of the signal $u(t)$ is chosen to be only 10% of the maximal deflection so that the maximal velocity $v_{lat}$ is not reduced. With the transfer function $H_2(s)$ and a maximal amplitude $y_{x,max}$ of $y_x(t)$ the amplitude $u_0$ is calculated in (17).

$$u_0 = y_{x,\max} \cdot \frac{K_D}{|H_2(i\omega_0)|} \tag{17}$$

Because $z_{rad}$ is located in the denominator of (17) its increase leads to a decreasing amplitude $u_0$. According to the described limit of 10%, $y_{x,max}$ is set to *R/20* with *R* being the radius of the beam.

# 5 PRACTICAL CONSIDERATIONS

The feedback controller is implemented in an analog design. To generate the signal $u(t)$, to measure the signal $v_1(t)$ and to adapt the gain of the controller, a microcontroller is used. The microcontroller is an ATmega128 and is programmed in C. The variable gain control is realized by digital potentiometers that are set by the microcontroller. Figure 12 shows the used circuit for one axis. The digital potentiometer $R_2$ offers 127 linearly arranged steps. The resistance can be adjusted between 1 kΩ and 50 kΩ.

The analog *PI* controller is built with standard components without complex serial or parallel circuits resulting in a time constant of $T_R$ = 132 µs and a adjustable gain between $K_R = 6.00 \cdot 10^{-3} \ldots 300 \cdot 10^{-3}$. The parameter $T_R$ remains constant even if the potentiometers change their value. Because of the discrete potentiometer positions there is an error between the optimal controller gain and the achieved controller gain. There are only integer positions $n_{int}$ available. To obtain an optimal value for $K_R$ the theoretical real number $n_{real}$ for the potentiometer position is calculated. Afterwards, $n_{real}$ is rounded down and up and the lower and the upper controller gains $K_l$ and $K_u$ are calculated. The controller gain with the minimal error in regard to the optimal value is chosen.

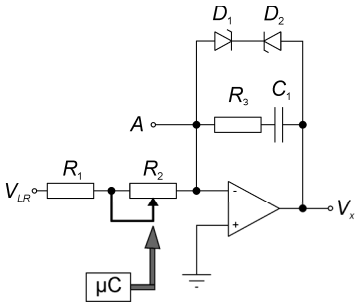



Figure 12: Microcontroller controlling a digital potentiometer and analog *PI* controller exemplarily shown for the x-axis.

After power up sequence and beam loss during operation, the laser beam searches the reflector in a defined area. This is done by deflecting the mirrors of the actuator without opening the feedback loop. The influence of the analog part is reduced and only the digital part controls the mirror deflection.

Figure 13 shows the operation principle. The digital potentiometer is set to its maximal value. So, the influence of the sensor signal to the input of the *PI* controller is weak. This is comparable to an opening of the feedback loop and the signal can be cross talked easily.

The microcontroller introduces a signal at the input of the *PI* controller. At the same time the reduced sensor signal of the diode acts as a disturbance variable. The output of the controller is digitalized and is compared to the set up variable $w_\alpha$. The microcontroller multiplies the gain $K_\mu$ with the deviation *e*. So, the plant with integrating behavior is controlled via a *P* controller which is a good combi-

nation (Merz & Jaschek 1996). The DA conversion at the output of the microcontroller is done via a pulse width modulation (PWM) with a frequency of 14.4 kHz and a low-pass filter with a cut-off frequency of 300 Hz (smoothing function).

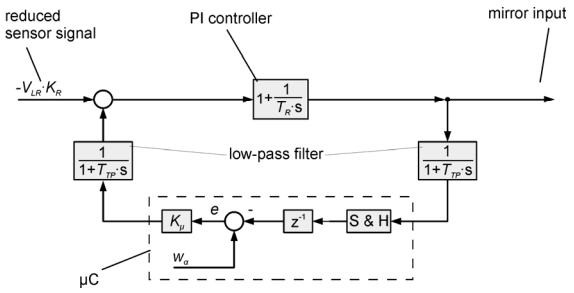



Figure 13: Operation principle for deflecting the mirrors in a defined way.

The output of the controller is digitized by the integrated AD converter of the microcontroller. To reduce alias effects a low-pass filter with a cut-off frequency of 300 Hz is used, too.

$K_\mu$ represents the total gain of the control loop and is set to $K_\mu = 41$. This is a tenth of the value of the stability limit. So, the safety margin is high enough to avoid instability.

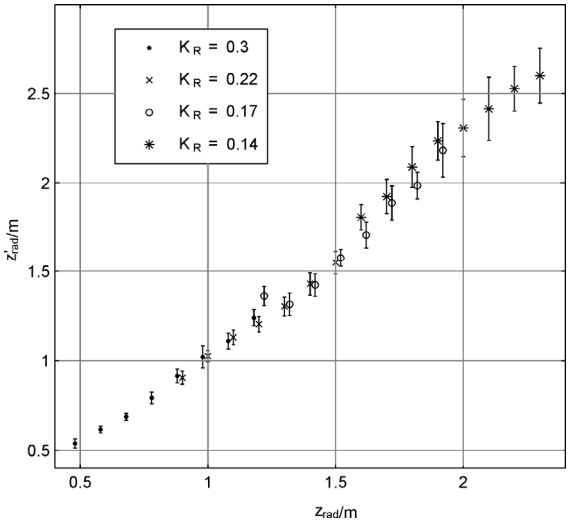



Figure 14: Estimated $z'_{rad}$ as a function of the real $z_{rad}$. Each point is averaged 20 times with its two-time standard deviation. To increase readability the measurements for $K_R = 0.3$ show a horizontal offset of -0.02 m and for $K_R = 0.17$ an offset of +0.02 m.

To estimate the distance $z_{rad}$ between the system and the retroreflector a signal is introduced in the x-axis (figure 11). To reduce the computing and programming complexity a square wave with a frequency of 200 Hz instead of a sinusoidal signal is used. An analog band pass filter with the same frequency is used for signal pre-processing. Only the basic frequency is considered in the signal analysis. Because the movement of the retroreflector has only few spectral components in the pass-band the gain can be increased without leaving the input range of the AD converter. So, a small value of $u_0$ is sufficient to detect the amplitude of $v_1(t)$ with a DTFT.

Figure 14 shows the distance estimation for different gains of the controller. To reduce the time effort only distances between 0.5 m and 2.3 m are measured. The estimated value shows a good accordance to the real values. The variance increases with the distance because of the reduced sensitivity of the measured voltage to the distance.

Further experiments have shown that the retroreflector can be moved with a maximal velocity of $v_{max} = 2.5$ m/s at unlimited acceleration and a beam offset of a quarter of the beam diameter. Distances between 0.5 m and 7.8 m (length of laboratory) were tested without tracking problems.

# 6 CONCLUSIONS

An analog *PI* controller with additional features was presented for the use in laser tracker systems. It is fast to detect a rapid beam movement and shows good control accuracy. Functions as gain scheduling and distance estimation are integrated in this hybrid design consisting of a digital and analog part.

# REFERENCES

Föllinger, O., 1994. *Regelungstechnik.* 8th ed. Heidelberg: Hüthig GmbH.

Merz, L. & Jaschek, H., 1996. *Grundkurs der Regelungstechnik.* 13th ed. München/Wien: R. Oldenburg Verlag.

Riemensperger, M. & Gottwald, R., 1990. Kern Smart 310 – Leica's Approach to High Precision 3D Coordinate Determination, In: *F. Löffler*, *2nd International Workshop on Accelerator Alignment*. Hamburg, Germany, 10-12 September 1990. pp. 183-200.

Schwarze, G., 1962. Bestimmung der regelungstechnischen Kennwerte von P-Gliedern aus der Übertragungsfunktion ohne Wendetangentenkonstruktion, *Messen, Steuern, Regeln,* 5, pp. 447-449.

# A DYNAMIC MODEL OF A BUOYANCY SYSTEM IN A WAVE ENERGY POWER PLANT

Tom S. Pedersen and Kirsten M. Nielsen
*Department of Automation and Control*
*Aalborg University, Fr. Bajersvej 7, Aalborg, Denmark*
*tom@es.aau.dk, kmn@es.aau.dk*

Keywords: Dynamic model, wave energy, simulation, buoyancy control, verification, renewable energy.

Abstract: A nonlinear dynamic model of the buoyancy system in a wave energy power plant is presented. The plant ("Wave Dragon") is a floating device using the potential energy in overtopping waves to produce power. A water reservoir is placed on top of the WD, and hydro turbines lead the water to the sea producing electrical power. Through air chambers it is possible to control the level, the trim and the heel of the WD. It is important to control the level (and trim, heel) of the WD in order to maximize the power production in proportion to the wave height, here the amount of overtopping water and the amount of potential energy is conflicting. Five separate air chambers, all open to the sea, makes the device float. The pressures in the air chambers may be individually controlled by an air fan through an array of valves. In order to make a model-based control system, this paper presents a model describing the dynamics from the air inlet to the level, trim and heel. The model is derived from first principles and is characterized by physical parameters. Results from validation of the model against plant data are presented.

## 1 INTRODUCTION

Renewable energy is an important issue due to the global warming problem and utilisation of wave power is one of the energy resources to be exploited.

The wave power system "Wave Dragon", on which this paper focuses, was invented by Erik Friis Madsen, Löwenmark and tested at Aalborg University and University of Cork. An EU based European consortium has been involved in the construction and implementation of a 1:4 scaled test site - 57x27 m wide and with a weight of 237 tonnes- which is placed in Nissum Bredning in Denmark. Large numbers of tests have been carried out during a two years operating period. One goal for energy production improvement is a better control of the Wave Dragon buoyancy.

Wave Dragon (WD) is an offshore wave energy converter of the overtopping type, a description is found in (Kofoed, 2006) and (W.D.Aps, 2006). The main structure consists of a ramp where the waves are overtopping and led to a reservoir (basin). Two reflectors are focusing the waves towards the ramp as seen on figure 1. WD is fastened to an anchor making it possible to turn the ramp towards the dominant wave direction.

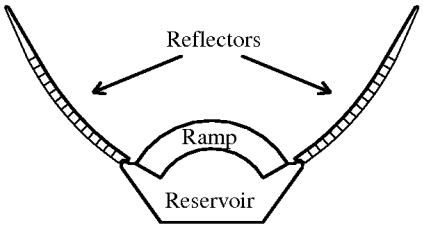



Figure 1: Main components of the Wave Dragon (Kofoed, 2006).

The WD use the potential energy of the waves, meaning that for a given wave type there exist an optimal level of the reservoir. As shown on figure 2 the reservoir water is led through a turbine.

The WD floats on open air chambers used to adjust the floating level. Control of the floating level is a part of optimizing the overtopping and a dynamic model for a model-based control system is the topic of this paper. It should be noted that the wave conditions are measured online and may be used as reference to the level control system.

First the wave dragon buoyancy system is presented. A dynamic model of the air supply system controlling the pressure in the air chambers is set up.

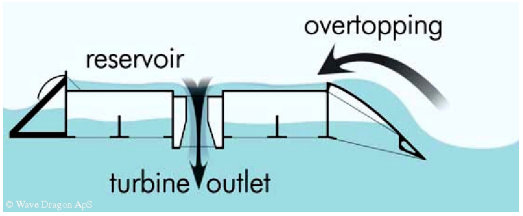



Figure 2: The basic principle of WD showing waves loading the reservoir via the ramp. (Kofoed, 2006).

The model parameters have been adjusted to the test site Wave Dragon. Finally the buoyancy model is verified by comparing simulation results and measurements from the test site Wave Dragon.

## 2 DYNAMIC MODEL OF THE BUOYANCY SYSTEM

The buoyancy system consists of five air chambers all open to the water surface. The five chambers are shown on figure 3 (21,22,23,24,25,26,27), (15,16,18,19,20), (3,4,5,10,11), (1,2,8,9,14) and (6,7,12,13,17). Furthermore 9 small chambers contain a constant amount of air.

The air pressures in the chambers are controlled by an air supply system using an on/off driven air fan and input/output valves to each chamber. The valves are operated as on/off valves and only one valve is allowed to be active at the time in order to prevent pressure equalizing in the chambers. A PWM scheme is in (Andersen, 2007) proposed to handle this problem. The air supply system model consist of two parts, one describing air inlet to the chamber and one describing air flow out of the chamber. In both models tube pressure drops are ignored. The inlet air mass flow, $m_{ai,}$ is given by

$$m_{ai} = -K_{pa}(p_c - p_a) + K_{pb}$$

which is an approximation to the fan characteristic. $p_c$ is the chamber pressure, $p_a$ is the inlet pressure to the fan (atmospheric pressure) and the two constants $K_{pa}$, $K_{pb}$ are from the fan data sheet.

Air outlet mass flow, $m_{ao,}$ is given by the Bernoulli equation:

$$m_{ao} = K_{vo}\sqrt{p_c - p_a}$$

where the constant $K_{vo}$ depend on the outlet tube dimension and the air density.

Each input/output valve pair is controlled by a signal u, where u=1 allows an airflow into the chamber, u=0 closes both valves and u=-1 open the outlet valve. This gives the air mass flow equation:

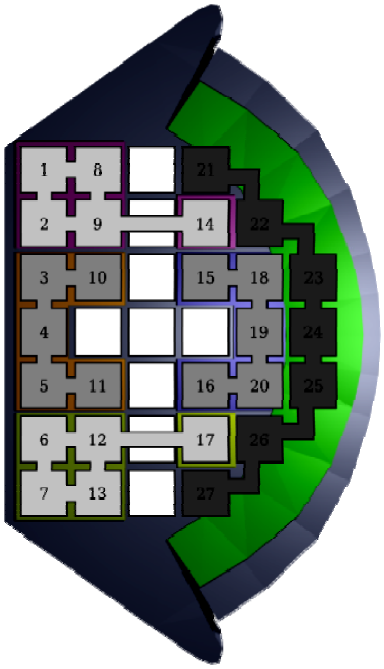



Figure 3: The air chambers in the Wave Dragon.

$$m_a = m_{ai}g_1(u) - m_{ao}g_2(u) \qquad (1)$$

where

$$g_1(u) = \begin{cases} 1 & if \quad u \geq 1 \\ 0 & if \quad u < 1 \end{cases}$$

$$g_2(u) = \begin{cases} 1 & if \quad u \leq -1 \\ 0 & if \quad u > -1 \end{cases}$$

Eq. (1) describes the air mass flow to the chamber and is valid when only one chamber is operated at a time.

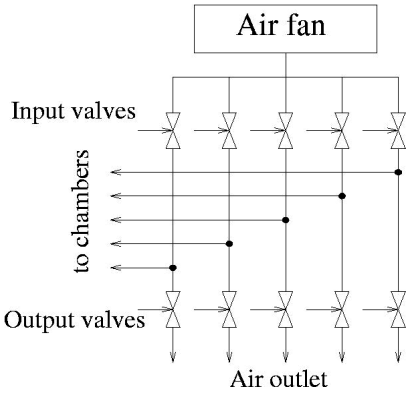



Figure 4: Fan/valve system.

# 3 ONE CHAMBER DYNAMIC MODEL

This section outlines the dynamic describing equations for a single chamber. The equations assumes that the air density is constant (variations could be included using the ideal gas equation), the chamber is only moving along a vertical axis perpendicular to the water surface, the cross section area is constant along this axis and there is only one rigid moving body. The model then consists of a mass balance equation describing the air in the chamber and a Newton equation describing the motion of the chamber.

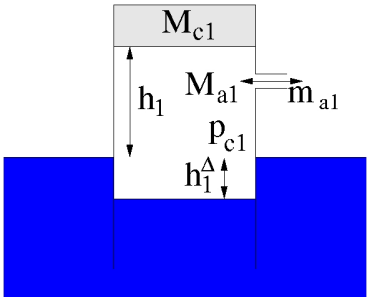



Figure 5: One chamber model variables.

The mass balance equation for the chamber is

$$\frac{dM_{a1}}{dt} = m_{a1} = A_1 \rho_a \frac{d(h_1^{\Delta} + h_1)}{dt} \qquad (2)$$

where $M_{a1}$ is the air mass in the chamber, $m_{a1}$ is the mass flow to the chamber, $A_1$ is the cross section area, $\rho_a$ is the air density, $h_1^{\Delta}$ is the distance from the chamber water surface to the ambient water surface and $h_1$ is the distance from the ambient water surface to the top of the chamber.

The pressure force from the chamber is assumed to equal the buoyancy force. This implies that the acceleration of the water volume in the chamber is small.

$$h_1^{\Delta} A_1 \rho_w g = p_{c1} A_1 \Rightarrow h_1^{\Delta} = \frac{1}{\rho_w g} p_{c1} \qquad (3)$$

where $\rho_w$ is the water density. Insertion of Eq. (2) in Eq. (3) gives

$$\frac{dp_{c1}}{dt} = \frac{g\rho_w}{A_1 \rho_a} m_{a1} - \rho_w g \frac{dh_1}{dt} \qquad (4)$$

The other main equation for describing the dynamics of the chamber is Newton's 2'th law used on the free body with the mass $M_{c1}$. Pressure forces, gravity and a friction force proportional with the chamber velocity is assumed to act on the body.

$$M_{c1} \frac{d^2 h_1}{dt^2} = (p_{c1} - p_a) A_1 - M_{c1} g - K_{f1} \frac{dh_1}{dt} \qquad (5)$$

In order to use an ODE-solver to simulate the one chamber model, the three states $x_1 = p_{c1}$, $x_2 = h_1$ and $x_3 = \frac{dh_1}{dt}$ may be selected resulting in the differential equation system

$$\begin{bmatrix} \frac{dx_1}{dt} \\ \frac{dx_2}{dt} \\ \frac{dx_3}{dt} \end{bmatrix} = \begin{bmatrix} -\rho_w g x_3 + \frac{g\rho_w}{A_1 \rho_a}((-K_{pa}(x_1 - p_a) + \\ K_{pb}) g_1(u) - K_{vo}\sqrt{x_1 - p_a}\, g_2(u)) \\ x_3 \\ (x_1 - p_a) A_1 - M_{c1} g - K_{f1} x_3 \end{bmatrix} \qquad (6)$$

# 4 FIVE-CHAMBER DYNAMIC MODEL

In this section a 5 chamber dynamic model is described. Although the WD has a complex geometry the model assumes that each of the five chambers may be regarded as a control volume with position independent internal variables and that each chambers pressure force actuate the WD bode in a single point. The model consist of 8 differential equations, 5 equations describing the pressures in the five chambers derived from mass balance equations, and 3 equations describing height, trim and heel derived from Newton's law.

The five chambers are placed in a "wave dragon" coordinate system {WD} with the coordinates $(\bar{x}_n, \bar{y}_n, \bar{z}_n)$. An inertial coordinate system {I} is placed in the water level as shown on the figure. The states in the model are the five chamber pressures ($p_{c1}$, $p_{c2}$, $p_{c3}$, $p_{c4}$, $p_{c5}$), the trim angle ($\theta$), the heel angle ($\gamma$) and the level (h) of the wave dragon.

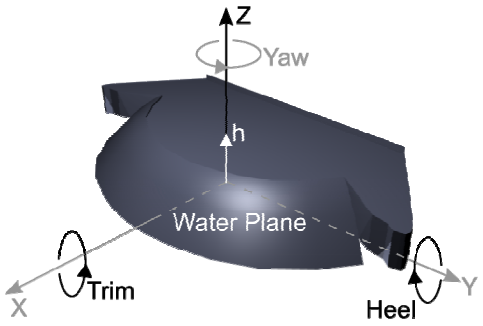



Figure 6: Wave dragon coordinate systems.

The mass balance for the n'th chamber is

$$\frac{dp_{cn}}{dt} = \frac{g\rho_w}{A_n\rho_a} m_{an} - \rho_w g \frac{dh_n}{dt}$$

$h_n$ is the height in the chamber and not a model state. Using a rotation matrix ${}^{I}_{WD}R(\theta,\gamma,\alpha)$ (see (Craig, 1989)), $h_n$ and the states are related through

$$h_n = -\bar{x}_n \sin(\gamma) + \bar{y}_n \cos(\gamma)\sin(\theta) + \bar{z}_n \cos(\gamma)\cos(\theta) + h$$

The time derivative of $h_n$ is a tedious equation but using the approximations $\cos(\theta)= 1$, $\cos(\gamma) = 1$, $\sin(\theta) = \theta$, $\sin(\gamma)= \gamma$ assuming small heal and trim angles (all $z_n$'s are 0) gives the simple relation

$$\frac{dh_n}{dt} = \bar{x}_n \frac{d\gamma}{dt} + \bar{y}_n \frac{d\theta}{dt} + \frac{dh}{dt}$$

which inserted gives

$$\frac{dp_{cn}}{dt} = \frac{g\rho_w}{A_n\rho_a} m_{an} - \rho_w g(\bar{x}_n \frac{d\lambda}{dt} + \bar{y}_n \frac{d\theta}{dt} + \frac{dh}{dt}) \quad (7)$$

Newton's 2'th law for the translational system is

$$\sum_{n=1}^{5} A_n p_{cn} + K_{flh}\frac{1}{h} - M_{wd}g - p_w A_w - K_f \frac{dh}{dt} = (M_{wd} + \frac{p_w A_w}{g})\frac{d^2h}{dt^2} \quad (8)$$

where the sum is chamber pressure forces acting on the wave dragon body. The second term is the forces from the air chambers containing a constant air mass (see figure 3). In the model these are approximated with two chambers. They may be modelled using the mass balance Eq. (4) but in order to keep the model order low they are modelled using a static balance. $K_{flh}$ may be found using the ideal gas law and the geometrics of the chambers. $M_{wd}$ is the total mass of the wave dragon. The term $p_w A_w$ represents the force from the water reservoir on top the Wave Dragon. It may be noted that the pressure $p_w$ is an input to the model.

The rotational trim equation is

$$\sum_{n=1}^{5} A_n p_{cn} \bar{y}_n - \frac{K_{fl\theta}}{h^2}\theta - K_\theta \frac{d\theta}{dt} = J_\theta \frac{d^2\theta}{dt^2} \quad (9)$$

where the sum is the torque from the chambers. The second term is the torque from the constant air mass chambers modelled as two symmetrical chambers. It should be noted that $K_{flh}$ as well as $K_{fl\theta}$ are very dependent on the total air mass in the chambers. The moment of inertia is calculated as a constant although it depends on the water level in a very complex manner. In the simulation a situation with low water level has been used.

The heel equation is

$$\sum_{n=1}^{5} A_n p_{cn} \bar{y}_n - K_{fl\gamma}\gamma - K_\gamma \frac{d\gamma}{dt} = J_\gamma \frac{d^2\gamma}{dt^2} \quad (10)$$

This linear second order equation captures the gross behaviour of the heel dynamics.

The total model now consist of eleven equations, equation (7) used 5 times for the five chambers giving the five states $x_1 = p_{c1}$, $x_2 = p_{c2}$, $x_3 = p_{c3}$, $x_4 = p_{c4}$, $x_5 = p_{c5}$, equation (8) with the states $x_6 = h$, $x_7 = \frac{dh}{dt}$ and the equation $\frac{dx_6}{dt} = x_7$ gives 2 equations, equation (9) with the states $x_8 = \theta$, $x_9 = \frac{d\theta}{dt}$ and the equation $\frac{dx_8}{dt} = x_9$ gives 2 equations, and finally equation (10) with the states $x_{10} = \gamma$, $x_{11} = \frac{d\gamma}{dt}$ and the equation $\frac{dx_{10}}{dt} = x_{11}$ gives 2 equations.

Inserting the states $(x_1, x_2, ..., x_{11})$ gives eleven nonlinear first order equations. The equations are solved using an ODE-solver.

# 5 SIMULATION

The model is tested using measured data from the wave dragon. The inputs to the model are the chamber pressures as well as the water level in the dragon represented by the pressure measurement $p_w$. The reason for not using the valve signals is that these were not recorded in the data acquisition equipment. As seen there is a good agreement between level in the model and the experimental data. (The agreements are found on the measurements from the same day. Measurements on different days are based on different initial pressures causing identification of slightly different model parameters) (This is also found on data measured on the same day.) Because the pressures in the constant air chambers are not measured these have been estimated from steady state observations.
The variations on the heel and trim angles are small. As seen on the figures the behaviour is captures by the model. All the data were recording prior to this project and not prepared for this study, unfortunately the WD run severely aground during the project meaning that controlled input signal could not be tested. Regardless of this the model performed well on the recorded data.

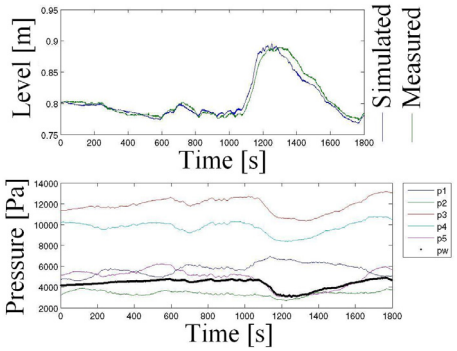



Figure 7: Level simulated and measured.

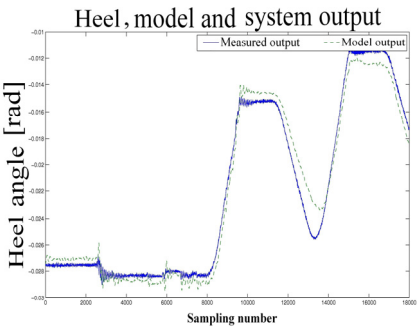



Figure 8: Heel simulated and measured (Andersen, 2007).

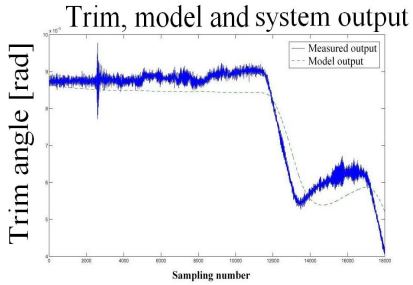



Figure 9: Trim simulated and measured (Andersen, 2007).

# 6 CONCLUSIONS

A nonlinear physical model with a complexity that is suitable for model based control has been presented. The model is partly based on physical parameters for the Wave Dragon and may be scaled to a future larger version. The model has four main equations, one describing the state of the air in a chamber, and three accounting for the level, trim and heel motion of the WD. The model has been validated against measured WD data, where it captures the gross behaviour of the Wave Dragon. In particular it describes the response of the level very well. The model does, however, have one serious deficiency because it does not capture the distribution and movement of water in the water reservoir. A comprehensive study of this is outside the scope of this paper.

# REFERENCES

Andersen, J., Hundsdahl, M.Y., Jensen, P.K.,
Vilbergsson, K.S., Vidarsson, O., Skagestad, R., 2007, *Control of Wave Dragon Buoyancy*, master thesis, Aalborg University.

Craig, J.J., 1989, *Introduction to Robotics,mechanics and control*, ISBN 0-201-09528-9.

Kofoed, J.P., Frigaard, P. Friis-Madsen, E. Sørensen, H.C. 2006, *Prototype testing of the wave energy converter wave dragon*, Renewable Energy 31 181-189.

W.D. Aps , 2006, *Wave Dragon – principles*, http://www.wavedragon.net

# ONE MODIFICATION OF THE CUSUM TEST FOR DETECTION EARLY STRUCTURAL CHANGES

Julia Bondarenko
*Department of Economic and Social Sciences, Helmut-Schmidt University Hamburg*
*University of the Federal Armed Forces Hamburg, Holstenhofweg 85, 22043 Hamburg, Germany*
*bonda@hsu-hh.de*

Keywords: CUSUM test for structural breaks, recursive residuals, power of test, mixture distribution, Monte Carlo simulation.

Abstract: Structural shifts in time series can occur as consequences of complex processes, arising in system. An ignorance of such structural changes can cause an associated regression model misspecification. In practice, early detection and response to outbreaks, causing the changes in a process, is highly important. The famous CUSUM test of Brown, Durbin & Evans, has a poor power in detecting the structural breaks in parameters occurring early (and also late) in the sample. In this paper, we propose CUSUM-similar test which, due to the transformation of recursive residuals forces the detection of temporal dependence structure in linear regression model and has a larger power for the early structural breaks. Here our interest centres on the detection of single breaks occurring in parameters of the linear model. Distribution and other probabilistic characteristics of the transformed residuals are provided, the boundaries for the new test are derived. The new test can be considered then as a complement to the standard CUSUM test.

## 1 INTRODUCTION

Structural shifts in time series can occur as consequences of complex processes, arising in system (Basseville and Nikiforov, 1993). An ignorance of such structural changes can cause an associated regression model misspecification. The evidence of the parameters instability in linear models can be detected by the number of corresponding diagnostic tests. Particularly, the famous and important of them include fluctuation tests, with the CUSUM and CUSUMQ tests of Brown, Durbin & Evans (BDE tests) (R. L. Brown and Evans, 1975), standing first on the list. These tests are easy for implementing and based on the calculating of the cumulative sums of recursive residuals (CUSUM) and the cumulative sums of the squares of recursive residuals (CUSUMQ) for regression. The CUSUM test is generally used to detect the systematic movements of parameters, whereas CUSUMQ test tends to capture the sudden, haphazard movements.

A number of extensions of the standard BDE CUSUM tests have been developed, like the CUSUM and CUSUM of squares tests using OLS (ordinary least squares) residuals instead of recursive residuals (Proberger and Kraemer, 1992), the CUSUM tests with lagged dependent variables in regression (W. Proberger and Alt, 1989), the CUSUM tests with non-stationary regressors (Inder and Hao, 1996), the MOSUM and MOSUM of squares tests putting emphasis on moving averages rather than cumulative sums (C.-S. J. Chu and Kuan, 1995a), the moving-estimates test (C.-S. J. Chu and Kuan, 1995b), etc.

It is a well-known fact, that, CUSUM tests, both with recursive and OLS residuals, have a poor power in detecting the structural breaks in parameters occurring early and late in the sample, as well as the ones orthogonal to the mean regressor (see (W. Kraemer and Alt, 1988); (Kraemer and Sonnberg, 1986), pp. 50-51; (Proberger and Kraemer, 1990); (Proberger and Kraemer, 1992)). A modification of recursive residuals CUSUM test, which is robust to the later problem, was proposed in (Luger, 2001). The reason for earlier problem is that CUSUM has no chance to cumulate for the such kind of breaks. In particular, some improving, due to alternative test boundaries developing, was suggested, in (Zeileis, 2000) for OLS CUSUM test.

In practice, early detection and response to outbreaks, causing the changes in a process, is highly important. In this paper, we propose a CUSUM-similar test which, due to the transformation of recursive residuals forces the detection of temporal dependence structure in linear regression model and has

a larger power for the early structural breaks. Here our interest centres on the detection of single breaks occurring in parameters of the linear model. Distribution and other probabilistic characteristics of the transformed residuals are also provided, the boundaries for the new test are derived. The new test can be considered then as a complement to the standard CUSUM test. The paper is organized as follows. Section 2 presents the BDE's original formulation of the CUSUM test, discusses its main features. Section 3 offers our modification of the CUSUM testing procedure. In the Section 4 we conduct comparison of two tests via Monte Carlo simulation study. Section 5 summarizes and concludes.

## 2 STANDARD CUSUM TEST

Consider the linear regression model

$$y_t = x_t'\beta_t + \varepsilon_t \ , t = 1,...,T, \tag{1}$$

where $y_t$ is the $t$th observation of the dependent variable, $x_t$ is the $k \times 1$ vector of covariates at time $t$, $\beta_t$ are unknown parameters, and $\varepsilon_t$ are independent and normally distributed with zero mean and variance $\sigma^2$. We are interested in testing for a discrete jump in at least one of the $\beta_t$ at the unknown time point. The null hypothesis can be formulated as

$$H_0 : \ \beta_t = \beta, t = 1,...,T, \tag{2}$$

with alternative $H_1$ that at time $T^*$ at least one of the $\beta_t$ changes its value:

$$H_1 : \ \beta_t = \beta, t = 1,...,T^* < T; \beta_t \neq \beta, t = T^*+1,...,T. \tag{3}$$

The CUSUM test suggested by Brown, Durbin and Evans (BDE) is a standard and commonly used diagnostic test for this kind of situations. BDE's CUSUM test uses the standardized residuals defined as

$$w_t = \frac{y_t - x_t'\widehat{\beta}_{t-1}}{\sqrt{1 + x_t \left(X_{t-1}'X_{t-1}\right)^{-1} x_t'}} \text{ for } t = k+1, k+2, ..., T, \tag{4}$$

where $\widehat{\beta}_{t-1}$ is the OLS-estimator of $\beta$ based on the first $t-1$ observations, $\widehat{\beta}_{t-1} = \left(X_{t-1}'X_{t-1}\right)^{-1} X_{t-1}Y_{t-1}$ , and $X_{t-1}$ and $Y_{t-1}$ are the $(t-1) \times k$ and $(t-1) \times 1$ matrices that obtain by stacking $x_s$ and $y_s$, respectively, for $s = 1, 2, ..., t-1$.

The advantage of working with $w_t$s defined by (4): it can be shown, that under $H_0$ (2), they are independent normally distributed with zero mean and variance $\sigma^2$. BDE CUSUM test is based on the cumulated sums of standardized residuals:

$$C_t = \frac{1}{\widehat{\sigma}} \sum_{s=k+1}^{t} w_s, \tag{5}$$

where $\widehat{\sigma}$ is OLS-estimate of $\sigma$, $\widehat{\sigma^2} = \frac{1}{T-k} \sum_{s=k+1}^{t} w_s^2$. We cannot obtain the explicit distribution of $C_t$. But, under null hypothesis of parameter stability, the expected value and variance of the statistic $C_t$ should be equal to zero and the number of normalized residuals being summed. Hence, the continuous Brownian motion process $Z_t$ can be considered as a good approximation of the discrete path of $C_t$. Under $H_0$ the sequence $C_t$ is thus a sequence of the approximately normal variables, where $E(C_t) = 0$, $Var(C_t) = t - k$, $Cov(C_t, C_r) = \min(t, r) - k$. Confidence bounds for the cumulated sums $C_t$ are then obtained by plotting the two straight lines connecting the points $k \pm a\sqrt{T-k}$ and $T \pm 3a\sqrt{T-k}$, where $a$ is a parameter depending on the $\alpha$ - significance level chosen, and is calculated on the basis of results for the Gaussian process. Namely, $a = 1.143$ for $\alpha = 0.01$; $a = 0.948$ for $\alpha = 0.05$; $a = 0.850$ for $\alpha = 0.1$ (see (R. L. Brown and Evans, 1975)).

Unfortunately, the CUSUM test suffers from low power, which decreases dramatically if the change point is close to the early beginning or to the end of the sample. In the next section we consider one modification of the CUSUM test based on the sconced cumulative sums.

## 3 MODIFICATION OF CUSUM TEST

Let's take into account the temporal structure of the residuals $w_t$, and construct a new sequence of random variables $u_t$, $t = k+1, ..., T$:

$$u_t = w_t + c\, w_t\, I\{w_t\, w_{t-1} > 0\}, \quad w_k = 0. \tag{6}$$

where, as before, $w_t \sim N\left(0, \sigma^2\right)$ and independent, $I\{\cdot\}$ is indicator function, and $c$ is some constant, $c > 0$, which can be considered as a penalty magnitude for upward or downward trends in CUSUM values. We will obtain now a distribution of the variables $u_t$.

**Theorem.** Let $w_t$ are independent identically distributed random variables, having a common symmetric about zero distribution, where $F_w(x)$ is a probability function, and $c > 0$. Then

$$F_{u_t}(x) = \frac{1}{2}\left[F_{w_t}(x) + F_{w_t}\left(\frac{x}{1+c}\right)\right]. \qquad (7)$$

*Proof.* We can perform $F_{u_t}(x)$ as a sum of the following probabilities:

$$\begin{aligned} F_{u_t}(x) \;=\; & P[u_t < x] = P[w_t + c\, w_t\, I\{w_t\, w_{t-1} > 0\} < x] = \\ & P[w_t + c\, w_t\ < x\,;\, w_t\, w_{t-1} > 0] + \\ & P[w_t < x\,;\, w_t\, w_{t-1} \le 0]. \end{aligned} \qquad (8)$$

For simplicity let's consider two cases:
1) $x \le 0$. Then (8) can be written as:
$P[w_t + c\, w_t\ < x\,;\, w_{t-1} < 0]$ $+$
$P[w_t \le x\,;\, w_{t-1} \ge 0] \;=\; P\left[w_t\ < \frac{x}{1+c}\,;\, w_{t-1} < 0\right]$ $+$
$P[w_t \le x\,;\, w_{t-1} \ge 0] \;=\; F_{w_t}\left(\frac{x}{1+c}\right) F_{w_t}(0)$ $+$
$F_{w_t}(x)\left(1 - F_{w_t}(0)\right) \;=\; F_{w_t}\left(\frac{x}{1+c}\right) F_{w_t}(0) + F_{w_t}(x) -$
$F_{w_t}(x) F_{w_t}(0) = \frac{1}{2}\left[F_{w_t}\left(\frac{x}{1+c}\right) + F_{w_t}(x)\right]$.
2) $x > 0$. (8) has a form:
$P[w_t + c\, w_t\ < 0\,; w_t\, w_{t-1} > 0]$ $+$
$P[0 \le w_t + c\, w_t < x\,;\, w_t\, w_{t-1} > 0] +$
$P[w_t\ < x\,; w_t\, w_{t-1} \le 0] \;=\; P[w_t < 0\,;\, w_{t-1} < 0]$ $+$
$P\left[0 \le w_t < \frac{x}{1+c}\,;\, w_{t-1} > 0\right]$ $+$
$P[w_t\ < x\,; w_t\, w_{t-1} \le 0] \;=\; F_{w_t}(0) F_{w_t}(0)$ $+$
$\left(F_{w_t}\left(\frac{x}{1+c}\right) - F_{w_t}(0)\right) F_{w_t}(0) +$
$P[w_t\ < 0\,; w_{t-1} \ge 0] + P[0 \le w_t < x\,;\, w_{t-1} < 0] \;=$
$F_w(0) F_w(0) +$
$\left(F_{w_t}\left(\frac{x}{1+c}\right) - F_{w_t}(0)\right) F_{w_t}(0) + F_{w_t}(0)\left(1 - F_{w_t}(0)\right) +$
$\left(F_{w_t}(x) - F_{w_t}(0)\right) F_{w_t}(0) = \frac{1}{4} + \frac{1}{2}\left(F_{w_t}\left(\frac{x}{1+c}\right) - \frac{1}{2}\right) +$
$\frac{1}{4} + \frac{1}{2}\left(F_{w_t}(x) - \frac{1}{2}\right) = \frac{1}{2}\left[F_{w_t}\left(\frac{x}{1+c}\right) + F_{w_t}(x)\right]$.

Thus, the CDF (PDF) of the random variables $u_t$ is a mixture of two normal distributions. It follows from the Theorem that $u_t$s have zero expectation, $E(u_t) = 0$, and variance $Var(u_t) = \frac{\sigma^2}{2}\left(1 + (1+c)^2\right)$. The joint CDF function obtained can be obtained by analogy with (7).

Covariance, $Cov(u_t, u_{t+1})$= $E[(u_t - E(u_t))(u_{t+1} - E(u_{t+1}))]$=$E(u_t\, u_{t+1})$, equals to zero. In addition, variables $u_t$s have zero skewness $s(u_t) = 0$, and kurtosis $\kappa(u_t) = \frac{6\left[1+(1+c)^4\right]}{\left(1+(1+c)^2\right)^2}$, revealing fatter tails.

The new statistic, which we will call "penalized" CUSUM (PCUSUM), can be written as following:

$$C_t^p = \frac{1}{\widehat{\sigma_p}} \sum_{s=k+1}^{t} u_s = \frac{1}{\widehat{\sigma^p}}\left(\sum_{s=k+1}^{t} w_s + c \sum_{s=k+1}^{t} w_s I\{w_s\, w_{s-1} > 0\}\right), \qquad (9)$$

where $\widehat{\sigma_p}^2 = \frac{1}{T-k} \sum_{s=k+1}^{t} u_s^2$. It is follows from the properties of $u_t$, that $E\left(C_t^p\right) = 0$ but $\widehat{\sigma_p} \ge \widehat{\sigma}$, thus, we expect that the PCUSUM test has a chance to cumulate well and to detect structural shifts only in very beginning of the sample.

Like the original CUSUM test (5), the modified test (9) is only an asymptotic test and has no any certain distribution. Obviously, under $c \to 0$, the sequence $C_{k+1}^p, ..., C_t^p$ may be approximated by the Brownian motion process mentioned in Section 1, and the bounds for cumulated sums $C_t^p$ are calculated then as the ones for $C_t$. But for larger values $c$ this can not be applied. Hence, we have here a subproblem of simulation of the presented test boundaries.

We will partially implement the techniques used in (Tanizaki, 1995) for confidence intervals calculation. The simulation algorithm may be carried out as follows. Let us generate $L$ replicates. In each replicate $i$, $i = 1, 2, ..., L$, we simulate $T - K$ random variables $u_{i,t}$, $t = k+1, k+2, ..., T$, pairwise dependent and uncorrelated, drawn from the mixture of two bivariate normal distributions (7). Then the cumulated sums, $s_{i,t}^p = \frac{1}{\widehat{\sigma^p}} \sum_{s=k+1}^{t} u_{i,s}$, are calculated. At the significance level $\alpha$ we have for statistic (9) that $P[L_{k+1}^p < C_{k+1}^p < U_{k+1}^p, ..., L_T^p < C_T^p < U_T^p] = 1 - \alpha$, where $L_t^p$ and $U_t^p$ are correspondingly the lower and upper bounds for the value $C_t^p$, and

$$P\left[L_{k+1}^p < C_{k+1}^p < U_{k+1}^p, ..., L_T^p < C_T^p < U_T^p\right] \neq \left(P\left[L_t^p < C_t^p < U_t^p\right]\right)^{T-k}$$

for any $t = k+1, ..., T$, since $C_t^p$ are not independent. Let assume that $P\left[L_t^p < C_t^p < U_t^p\right] = 1 - \alpha_p$, and denote $\alpha = f(\alpha_p)$, where $f$ is some unknown function (in the case of independence one has $f(x) = 1 - x^{T-k}$), which we will obtain by simulation. Applying the Newton-Raphson algorithm, we calculate our $\alpha_p$ following the scheme:

$$\alpha_p^{(j)} = \alpha_p^{(j-1)} + d^{(j)}\left(\alpha - f\left(\alpha_p^{(j-1)}\right)\right),$$

where $j$ is iteration number, $d^{(j)} = \delta d^{(j-1)}$ with $\delta = 0.5$ and $d^{(0)} = 1$, $\alpha_p^{(0)} = \alpha$. The convergence criterion is $\left|\alpha_p^{(j)} - \alpha_p^{(j-1)}\right| < 0.0001$. Repeat, that the function $f\left(\alpha_p^{(j)}\right)$ is derived at $j$th iteration as following: $f\left(\alpha_p^{(j)}\right)$ equals to the number of sequences $\left\{ s_{i,k+1}^p, ..., s_{i,T}^p \right\}_{i=1}^{L}$ within intervals $\left(L_{i,k+1}^p, U_{i,k+1}^p\right)^{(j)}$, ..., $\left(L_{i,T}^p, U_{i,T}^p\right)^{(j)}$ divided by $L$, where intervals $\left(L_{i,t}^p, U_{i,t}^p\right)^{(j)}$ are obtained for the value $\alpha_p^{(j)}$. Note, that unlike the standard CUSUM statistic, which has symmetric lower and upper boundaries, we don't claim here $L_t^p = -U_t^p$ for PCUSUM statistic. To be fair to the standard

CUSUM, we use a linear regression to obtain the linear boundaries from the curved ones: $\widetilde{L}_t^p = a_{1L} + a_{2L}\,t$ and $\widetilde{U}_t^p = a_{1U} + a_{2U}\,t$, where $a_{1L}$, $a_{1U}$ are intercepts, $a_{2L}$, $a_{2U}$ are slopes, $\widetilde{L}_t^p$, $\widetilde{U}_t^p$ are the best fitting lines.

# 4 MONTE CARLO STUDY

In this section we present a Monte Carlo study of the new CUSUM test in comparison with the classical one. In order to see the performance of $C_t^p$, we consider the model (1), where the matrix of independent variables $x_t$ has the following design:

$$X^{MC} = \left\{ \left[1, (-1)^t\right]' \right\}_{t=1}^{T}, \qquad (10)$$

which was also used in simulation study in (Inder and Hao, 1996) and (Kraemer and Sonnberg, 1986). Values $\varepsilon_t$, as before, are independent normal, generated with parameters 0 and 1. We simulate our responses $y_t$ for three sample sizes, $T = 20$ (small sample), $T = 50$ (medium sample) and $T = 500$ (large sample), with $\beta = [10, 2]'$ under the null hypothesis of parameters constancy. The significance level $\alpha = 0.05$, the computed empirical levels $\alpha_p$ for samples under $c = \{0.2;\ 0.5;\ 1.5\}$ are correspondingly $\alpha_p = \{0.0075; 0.0049; 0.0054\}$ ($T = 20$), $\alpha_p = \{0.0055; 0.0036; 0.0029\}$ ($T = 50$), $\alpha_p = \{0.0011; 0.0013; 0.0016\}$ ($T = 500$).

Empirical (actual) test sizes, based on the simulated data, at the nominal size of 5%, and $c = \{0.2; 0.5; 1.5\}$ in PCUSUM test, are presented in the Table 1. Generally estimated sizes, calculated as rejection rates under null hypothesis, with Monte-Carlo replications number $N = 5000$, are either below the nominal size for CUSUM test, resulting in more "liberal" test, or almost equal to the nominal size. However, the empirical sizes for PCUSUM test under $c = 1.5$ are larger than 0.05, declaring the more "sensitive" test.

Table 1: Empirical Sizes of the Tests, $\alpha = 0.05$.

| **Test** | $T = 20$ | $T = 50$ | $T = 500$ |
|---|---|---|---|
| CUSUM | 0.0154 | 0.0276 | 0.0458 |
| PCUSUM, $c = 0.2$ | 0.0338 | 0.0453 | 0.0476 |
| PCUSUM, $c = 0.5$ | 0.0512 | 0.0510 | 0.0444 |
| PCUSUM, $c = 1.5$ | 0.0686 | 0.0586 | 0.0646 |

Now, we introduce at time $T^* = [\lambda T]$, where $\lambda$ can take any values between 0 and 1, some structural shift. Let's consider a single structural shift in parameters $\beta$ is given by $\Delta\beta = \frac{b_0}{\sqrt{T}}\,[\cos\phi,\ \sin\phi]'$, where $\phi$ is the angle between $\Delta\beta$ and mean regressor $r = \left[\frac{1}{T}\sum_{t=1}^{T} x_{1t}, \frac{1}{T}\sum_{t=1}^{T} x_{2t}\right]' = [1, 0]'$, $b_0$ is a constant determining the intensity of the shift , $\|\Delta\beta\| = \frac{|b_0|}{\sqrt{T}}$. We will take a number of different values of $b_0$ and $\phi$, namely, $b_0 = \{-12;\ -8;\ -6;\ -3;\ 3;\ 6;\ 8;\ 12\}$ (positive and negative values, as we don't claim the boundaries of PCUSUM to be symmetric) and $\phi = \{0\}$. In other words, we are testing two hypotheses: $H_0$ : parameters are constant for $1 \le t \le T$ against $H_1$ : parameters have two different constant values, for $1 \le t < T^*$ and $T^* \le t \le T$.

It is possible to show that, as we have expected, the linear boundaries for PCUSUM test become narrower than CUSUM test boundaries only at the beginning of the sample. Hence, it makes sense to choose $\lambda$ corresponding to the structural changes at the beginning of the sample, $\lambda = 0.3$, for example. Empirical power was calculated as a probability that the test statistic under the alternative hypothesis exceeded the significance threshold calculated from the distribution under the null hypothesis (a frequency of the null hypothesis rejection under the alternative hypothesis). The obtained power plots, under $N = 1000$, $\lambda = 0.3$, $\alpha = 5\%$, $c \in \{0.2;\ 0.5;\ 1.5\}$ for $T = 20$, $T = 50$ and $T = 500$ are presented on the Figures 1, 2 and 3, correspondingly.

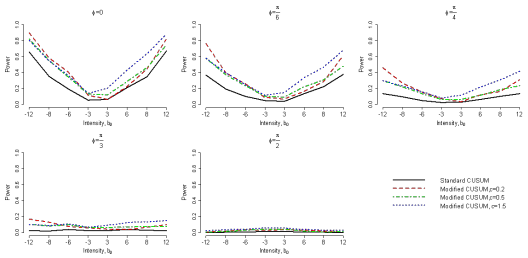



Figure 1: Power plots under $T = 20$, $\lambda = 0.3$.

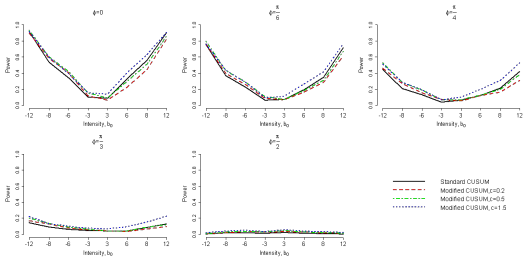



Figure 2: Power plots under $T = 50$, $\lambda = 0.3$.

The obtained simulation results for the small sample with $T = 20$ reveal that the PCUSUM outperforms the CUSUM. For small values of shift intensity $b_0$ and for $\phi = 90$ this superiority is quite insufficient, for all samples. For the medium sample $T = 50$, PCUSUM with $c = 1.5$ has higher power ev-

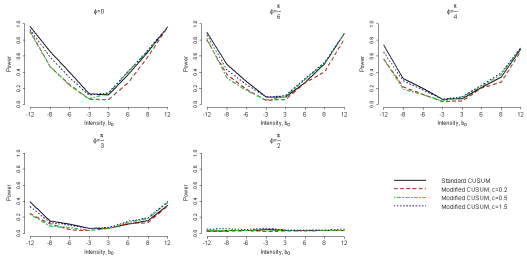



Figure 3: Power plots under $T = 500$, $\lambda = 0.3$.

erywhere, but PCUSUM with $c = 0.5$ and $c = 0.2$ - only for negative values of $b_0$. In big sample with $T = 500$ an advantage is exhibited only by PCUSUM with $c = 1.5$, for positive $b_0$s. By virtue of the boundaries non-symmetry, the results are different for positive and negative values of $b_0$ also among PCUSUM tests: for example, for $T = 20$ PCUSUM with larger parameter $c$ outperforms PCUSUM with smaller one for positive $b_0$s, but for negative $b_0$s PCUSUM with smaller value of $c$ remains more effective.

## 5 CONCLUSIONS

In this paper we proposed a modified, based on the penalized residuals, version of the standard BDE CUSUM test for single structural breaks in parameters of linear regression. The new test, PCUSUM, is recommended as a complement to the standard CUSUM test for better detecting the structural shifts occurring early in the samples. Simulation results have shown, that the modified CUSUM test has the better chance to cumulate parameter breaks, occurred at the beginning of the sample.

The subjects of eventual future research are:

- closer examination of properties of the PCUSUM test boundaries, their comparison with derived curved boundaries of the standard CUSUM test;
- adoption of the modified residuals (6) by CUSUM of squares test (CUSUMQ) for testing structural changes in variance (serial correlation and heteroscedasticity);
- appropriate transformation of the proposed new test for both early and late structural breaks (inclusion of the more lagged residuals in (6), etc).

## REFERENCES

Basseville, M. and Nikiforov, I. (1993). *Detection of Abrupt Changes - Theory and Application*. Prentice Hall, Englewood Cliffs, NJ.

C.-S. J. Chu, K. H. and Kuan, C.-M. (1995a). Mosum tests for parameter constancy. *Biometrika*, 82:603–617.

C.-S. J. Chu, K. H. and Kuan, C.-M. (1995b). The moving-estimates test for parameter stability. *Econometric Theory*, 11:669–720.

Inder, B. and Hao, K. (1996). A new test for structural change. *Empirical Economics*, 21:475–482.

Kraemer, W. and Sonnberg, H. (1986). *The Linear Regression Model under Test*. Physica-Verlag, Heidelberg.

Luger, R. (2001). A modified cusum test for orthogonal structural changes. *Economic Letters*, 73:301–306.

Proberger, W. and Kraemer, W. (1990). The local power of the cusum and cusum of squares tests. *Econometric Theory*, 6:335–347.

Proberger, W. and Kraemer, W. (1992). The cusum test with ols residuals. *Econometrica*, 60:271–285.

R. L. Brown, J. D. and Evans, J. M. (1975). Techniques for testing the constancy of regression relationships over time. *Journal of Royal Statistical Society B*, 27:149–163.

Tanizaki, H. (1995). Asymptotically exact confidence intervals of cusum and cusumq tests: a numerical derivations using simulation technique. *Communications in Statistics*, 24:1019–1036.

W. Kraemer, W. P. and Alt, R. (1988). Testing for structural change in dynamic models. *Econometrica*, 56:1355–1369.

W. Proberger, W. K. and Alt, R. (1989). A modification of the cusum test in the linear regression model with lagged dependent variables. *Empirical Economics*, 14:65–75.

Zeileis, A. (2000). *p*-values and alternative boundaries for cusum tests. *SFB Adaptive Information Systems and Modelling in Economics and Management Science*, Working Paper 78.

# INSTRUMENTING BOMB DISPOSAL SUITS WITH WIRELESS SENSOR NETWORKS

John Kemp, Elena I. Gaura and James Brusey
*Cogent Computing Applied Research Centre, Coventry University, Priory St, Coventry, CV1 5FB, U.K.*
{*kempj, e.gaura, j.brusey*}*@coventry.ac.uk*

Keywords: Body sensor networks, first responders, actuation.

Abstract: Bomb disposal suits contain a large amount of padding and armour to protect the wearer's vital organs in the case of explosion. The combination of the heavy (roughly 40kg) suit, physical exertion, and the environment in which these suits are worn can cause the wearer's temperature to rise to uncomfortable and potentially dangerous levels during missions. This paper reports on the development of a wearable wireless sensing system suitable for deployment in such manned bomb disposal missions. In its final form, the system will be capable of making in-network autonomous decisions related to the actuation of cooling within the suit, in order to increase the comfort of the wearer. In addition, it will allow an external observer to remotely monitor the health and comfort of the operative. Laboratory experiments with the instrumented suit show how skin temperature varies differently for different skin sites, motivating the need for multiple, distributed sensing. The need for timely application of in-suit cooling is also shown, as well as the importance of monitoring the overall health of the wearer of the suit.

## 1 INTRODUCTION

The monitoring of hazardous environments, along with the people working within them, is an area which lends itself to the use of wireless and body sensor networks (WSNs and BSNs). The field is rich with potential WSN applications in detecting hazards, providing feedback to remote observers and other critical tasks that can increase the safety and benefit the overall working conditions of people operating in these environments. This paper reports the work towards the development of a wireless body sensor network for the protective suits worn in bomb disposal missions.

A typical bomb disposal mission will initially involve investigating the site using a remote controlled robot, and if possible, disarming the bomb remotely. Sometimes, however, it is necessary for a human bomb disposal expert to disarm the device. For this, the expert will put on a protective suit and helmet (as shown in figure 1), pick up a tool box of equipment, and walk the 100 or so metres to the site. To reach the bomb's location, it may be necessary to climb stairs, crawl through passageways, or even lie down.

The environment where the suit is used, such as the hot climate of the Middle East, plays an important role in the design of the protective suit. One of the UK manufacturers of such suits has identified the problem of the suit wearer becoming uncomfortably hot and, in the worst case, suffering heat exhaustion. They have attempted to address this by installing an in-suit cooling system based on a dry-ice pack and a fan that cycles air through the pack and blows cooled air onto the wearer's back and into the helmet. The cooling system has a variable control thus both allowing the airflow to be adjusted for comfort and also allowing the life of the batteries that power the fan to be extended, as they would only provide sufficient power for part of the mission otherwise. The problem with this cooling approach, though, is that the bomb disposal expert has other critical concerns during the mission and either does not bother to put the fan on or tends to set it to maximum airflow from the beginning

Figure 1: Explosive Ordinance Disposal (EOD)Suit.

of the mission.

To address the above problems, this work proposes embedding into the suit a body sensor network that aims to:

- sense the temperature of the skin of various parts of the body, in order to assess overall comfort, and
- adjust the cooling dynamically to both remove the need for human intervention, and also to prolong battery life.

The prolonging of battery life is intended to provide cooling over the whole mission duration (compared to the partial coverage provided currently) rather than increasing the mission duration itself.

A secondary goal of this work is to help the manufacturer better understand how the suit material and design choices are affecting the wearer's thermal comfort during use. Finally, the prototype presented here has been designed such as to allow easy integration of additional sensors, such as accelerometers to monitor posture, heart rate monitors, and $CO_2$ sensing within the helmet.

The paper is organised as follows: Section 2 examines related work, focusing in particular on body sensor networks and research relating to instrumenting first responders (such as police, fire services etc). Section 3 describes the system design and architecture developed for the prototype system produced to date. Section 4 contains an evaluation of the prototype. Finally the paper concludes with some observations based on the work so far and outlines future work.

# 2 RELATED WORK

The work reported in this paper is most closely aligned with respect to the instrumentation design and implementation with the field of Body Sensor Networks. This is a sub-area of Wireless Sensor Networks that makes use of a combination of wireless and miniaturised sensor technologies to monitor the human body. The scope of present BSN approaches is patient care. Such systems are either designed to focus on capturing the evolution of a particular physiological parameter and ensuring that alarms are generated when parameters stray outside a safe range (Keoh et al., 2007), or aimed to provide general monitoring solutions for patient status within a hospital or similar environment (Shnayder et al., 2005). In comparison, the work presented here is concerned with increased safety and comfort of human subjects in constrained environments through integrating sensing, actuation, and autonomous decision making. In this context, wireless sensor technology is used as an enabler for the necessary detailed measurement of physiological parameters. 5A This work shares some of the design space of BSN in terms of the type of physiological parameters sensed and the wearability requirements of the implemented system. On the other hand, given that the application is within the safety critical domain, the work here also shares some common characteristics with the area of instrumenting and monitoring first responders. In this section, samples of BSN platforms are reviewed together with commercial instances of first responder monitoring and prior, motivating, physiological findings about the EOD suit.

## 2.1 Body Sensor Networks—Platforms

BSN based systems are often more constrained than ordinary embedded systems. These constraints are mainly in terms of power, size and weight. Power is restricted because mains AC power is not available. Furthermore, size and weight restrictions limit the battery supplies that can be used. Size and weight must be limited because large and heavy devices would be cumbersome, uncomfortable, and in applications such as the one described here, an unnecessary distraction.

In response to the above, some of the BSN systems designed and implemented by research groups integrate within the nodes an appropriate central processing unit, memory and radio transceiver as a single custom chip. An example here is the MITes platform (for monitoring movement of human subjects) developed by (Tapia et al., 2004), which is based around the Nordic VLSI Semiconductors nRF24E1 chip. This chip integrates a radio transceiver and an Intel 8051 based processor core that runs at 16MHz and provides a nine channel 12-bit ADC and various other interfaces, such as SPI (serial peripheral interface) and GPIO (general purpose I/O). This approach is efficient in terms of size and weight due to the integration of several functions into one chip, but has limited generality as it can not be easily adapted for new applications.

Another, more popular design option is to use off-the-shelf components. There is a trade off made between processing and storage capabilities and the size and power consumption of the devices. This means that the devices selected would likely be considered severely under-powered in other systems (often including 16- or even 8-bit processors) and have small amounts of memory (in the order of tens or hundreds of kilobytes). For instance, the Texas Instruments MSP430F149 micro-controller has been used for several systems including those developed by (Lo and

Yang, 2005) and (Jovanov et al., 2001). This is a 16-bit processor running at 8MHz incorporating 60KB of flash memory and 2KB of RAM and provides interfacing opportunities via 48 GPIO lines and a 12-bit ADC. The system developed by Lo and Yang used ECG sensors, accelerometers, and a temperature sensor to monitor patient health. The system developed by Jovanov et al., was used for monitoring the elderly and those undergoing physiotherapy.

Other systems expand upon commercial devices such as the Mica2 and MicaZ motes developed at the University of California, Berkeley, or Intel's Imote platform. This approach often has a disadvantage in that the basic platform is generic, and may not directly provide the facilities required for the specific BSN project. Such commercial platforms are also often larger and heavier than custom developed platforms as they are required to be general purpose in order to achieve any commercial success. The MicaZ mote uses the Atmega128L, an 8-bit processor running at 8MHz and featuring 128KB of flash memory to which an additional 512KB is added externally on the mote itself. A 10-bit ADC, UART and I2C bus are also available. (Gao et al., 2005) developed a system based around the this mote, adding various sensors and supporting devices to allow patient tagging and monitoring in an emergency response environment. (Walker et al., 2006) present a blood pressure monitoring system based on the MicaZ platform. In that work, a commercial blood pressure monitoring device is connected to the MicaZ via a serial interface.

## 2.2 Instrumenting First Responders

The best fit example of a commercial product designed for the purpose of monitoring personnel carrying out missions in dangerous environments is the VivoResponder by (Vivometrics, 2007). VivoResponder is based upon an earlier product called the LifeShirt and is aimed at personnel engaged in firefighting and hazardous materials training or emergency response, industrial clean-ups using protective gear, and biohazard-related occupational work. The VivoResponder is supplied in three parts: a lightweight, machine washable chest strap with embedded sensors; a data receiver; and, VivoCommand software for monitoring and data analysis. The sensors embedded in the chest strap monitor the subject's breathing rate, heart rate, activity level, posture, and single point skin temperature.

Monitoring of the subject's breathing is performed using a method called inductive plethysmography, where breathing patterns are monitored by passing a low voltage electrical current through a series of contact points around the subject's ribcage and abdomen. Monitoring of the subject's heart rate is performed via an ECG.

The VivoCommand software, provided with the device, displays the gathered data from the chest strap in real-time on a remote PC. The parameters are updated every second along with 30-second average trends. The parameters are displayed with colour coding intended to allow quick assessment of the status of up to 25 monitored personnel simultaneously. Baseline readings can be set individually per monitored person.

The work developed here differs in intent: the aim here is to provide a detailed thermal assessment based on sensors integrated into the protective suit and deliver remotely abstracted comfort information.

## 2.3 Other Work on EOD Suits

Working from a physiological perspective, (Thake and Price, 2007) have investigated the thermal strain of a subject when wearing EOD suits in hot environments. The work looks at quantifying the level of strain by assessing how hot and tired the suit wearer feels whilst wearing various combinations of suit components. An "activity" regime was developed for the assessment based on the types of activities that a bomb disposal technician would undergo during a mission and included walking on a treadmill, unloading and loading weights from a rucksack, crawling and searching activity, arm cranking and cognitive tests. Aspects of hand-eye coordination and psychological performance were also assessed. The investigations have demonstrated a large increase in physiological strain when wearing the EOD suit, though benefits have been shown when the ambient air is cooled for the suit ventilation purpose and lighter-weight trousers are worn.

It is indeed these types of studies, together with the user requests, that prompted the development of the detailed physiological monitoring system presented in this paper. The activity regimes described by Thake and Price were used in the experiments presented in this paper to allow validation of findings.

# 3 SYSTEM DESIGN AND ARCHITECTURE

The main part of the prototype system is designed following a sense-model-decide-act architecture as shown in figure 2. The environment within the suit is sensed in terms of temperature; sensed data is integrated into a model representing the thermal state

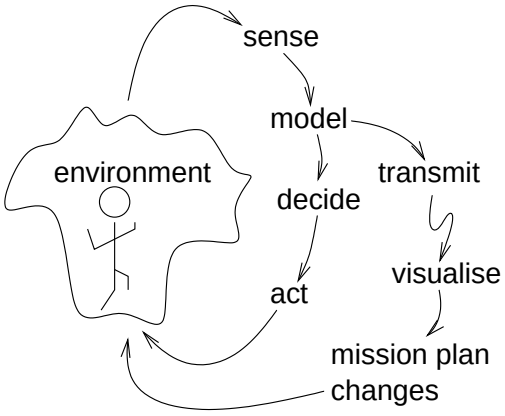



Figure 2: Conceptual design of prototype system.

of the wearer; a decision is made about how to adjust the cooling system based on the thermal state; finally, the determined action is transmitted to the fan speed controller. In addition to this basic architecture, the system also transmits inferred state values for the purpose of remote, on-line, visualisation of the thermal state of the wearer. From this visualisation, the operator can assess how different parts of the mission, or different actions being taken by the suit wearer are affecting their thermal state and hence assess the wearer's fitness for the mission. (It is expected that such information, collected during field trials and real missions, might lead to changes to future mission planning or to changes in the design of the suit.) In summary, the prototype system can be seen as being composed of two control loops: one giving rapid feedback to autonomously adjust cooling; the other, longer term one, providing support for an iterative design process in terms of both the mission use and construction of the suit.

The prototype design consists of a number of hardware components, including a remote monitoring station, two processing nodes, one actuation node, 12 temperature sensors, and the cooling system. The connection between these components is shown in figure 3. The processing nodes, actuation nodes, and remote monitoring point form a wireless network. Each processing node is wired to several sensor packages via an I2C bus. Although it would be possible to integrate all sensor packages used in this prototype into a single processing / actuation node, using separate processing nodes allows the helmet, jacket, and trousers to be kept separate with no wires running between them. This is essential for ensuring that the product remains easy to use and transparent to the wearer.

The system components and their functionality are described in the remainder of this section.

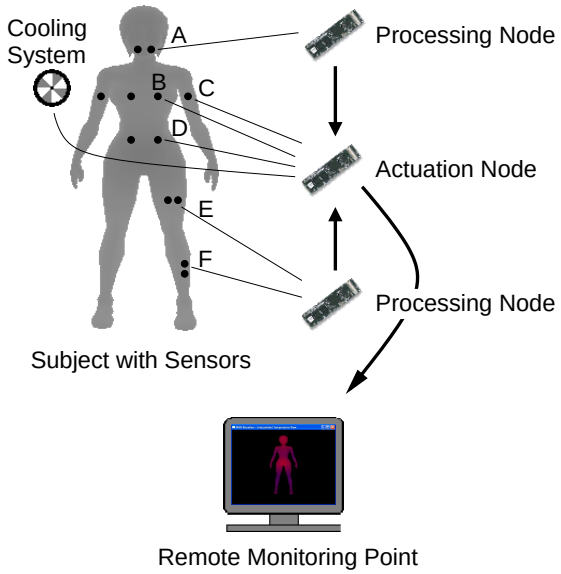



Figure 3: Prototype system hardware components and sensor positioning (A – neck, B – chest, C – bicep, D – abdomen, E – thigh, F – calf).

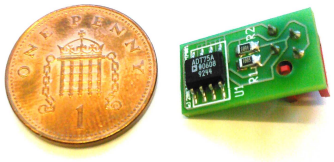


Figure 4: Sensor package, which is based on an ADT75A chip.

## 3.1 Sensor Packages and Sensor Positioning

The prototype system discussed here uses twelve sensor packages based on Analog Devices ADT75A temperature sensor ICs (shown in figure 4). This device has the advantage that it contains the sensor, ADC, and bus interface in a single package. Temperature values are transmitted as 12 bits, which causes rounding to within 1/8°C. The sensor packages are connected to only two nodes in the current version: one actuation node and one processing node.

The sensor packages were positioned around various parts of the body roughly following the standard positioning used for skin sensors as used by (Thake and Price, 2007), which is a subset of the locations discussed by (Shanks, 1975). These were: lateral calf muscle, front of thigh (or quadruceps), abdomen, chest, biceps, and neck, as indicated in figure 3. Given that temperatures are known to be symmetrical between left and right sides in healthy people (Silberstein et al., 1975), sensors have been placed on a single side. Two sensor packages were used per skin site.

This arrangement enables individual data validation.

## 3.2 Processing and Actuation Nodes

### 3.2.1 Construction

There are a variety of available embedded platforms for sensing and control applications. The hardware choice decisions for the prototype system here were based on the available platforms' processing power, external interfaces, ease of software development, and size.

Gumstix Connex 400xm-bt boards were selected as the main processing platform. Although not as popular as Mica2 motes, they are becoming more prevalent (see (Keoh et al., 2007) for an example). These devices offer more processing power and memory (in terms of both RAM and flash) than many similarly sized platforms. The Connex includes an Intel XScale PXA255 400MHz processor, 16MB of flash memory, 64MB of RAM, a Bluetooth controller and antenna, and 60-pin and 92-pin connectors for expansion boards. There are no on-board sensors provided. The sensor packages connect to the Connex board via an expansion board, designed in-house.

The prototype system exploits the following capabilities offered by the Gumstix Connex device: Bluetooth communications to transmit data between nodes; I2C bus interface for the attachment of sensor packages; real-time data modelling and decision-making; and, a small form factor, which enables convenient mounting on or around a subject's body.

### 3.2.2 Functionality

In the current revision of the prototype, the actuation and processing nodes only transmit data back to the remote monitoring station, upon filtering outlying values. The longer term view is for the processing and actuation nodes to perform in-network modelling of the suit wearer's comfort through collaborative behaviour. Comfort modelling would firstly involve production of a thermal sensation model within the network. This will be followed by integration of supplementary physiological and contextual sensing performed by expanded sensor packages. Work to date in thermal sensation modelling and its integration within the processing and actuation nodes is reported in a separate paper. The actuation node will eventually be used to perform decision making on the basis of the wearer's comfort and act by controlling the fan speed.

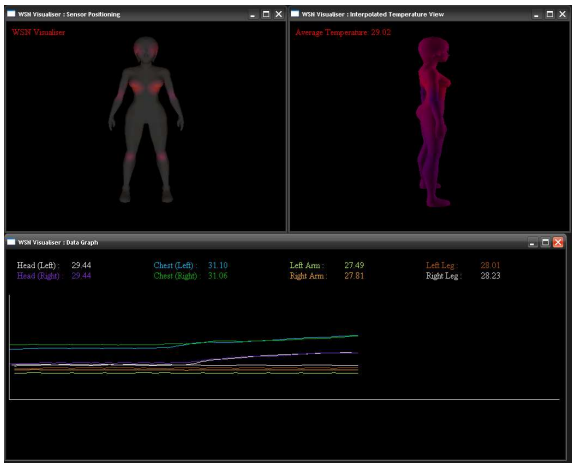



Figure 5: A snapshot of the remote monitoring component.

## 3.3 Remote Monitoring

The remote monitoring component of this system allows an external observer to monitor both the instrumentation system (to ensure that trustworthy information is being recorded) and the bomb disposal technician during a mission (which is the main function of the instrumentation). The remote monitoring component displays the health and comfort information and provides alerts to the remote observer if physiological parameters fall outside safe ranges or the wearer is shown to be significantly uncomfortable. A snapshot of the remote monitoring component is shown in figure 5. Currently the remote monitor displays skin site temperature data and a rotating, suggestive, 3-D interpolated model of skin temperatures. Cool to hot zones are displayed dynamically through a range of colours, from blue to red.

# 4 PROTOTYPE EVALUATION

## 4.1 Experimental Setup

The prototype instrumentation system was evaluated through laboratory experiments that attempt to reproduce typical bomb disposal mission situations by having the subject undertake a series of activities and tasks, as discussed in section 2.3. The experiments begin with sensors being attached to the subject, followed by suiting-up. The upper body sensors are integrated into the clothing and thus easier to attach, whilst lower body sensors are attached with PVC tape. The subject wore the outer shell of the bomb disposal suit including the jacket and trouser segments in addition to armour plating and the helmet.

Figure 6: First activity: walking at 4km/h.

Figure 7: Second activity: kneeling while removing weights from a sack. The wired-in data logger can be seen taped onto the subject's lower back.

The subject then undertakes an activity regime composed of: (1) walking (3 minutes) (see figure 6); (2) kneeling while putting weights into and out of a rucksack (2 minutes) (see figure 7); (3) crawling (2 minutes); (4) arm exercise (4 minutes); (5) sitting (3 minutes); (6) standing (1 minute). Temperature data is collected both via the prototype wireless system and via a wired-in data logger. Data was gathered during two consecutive runs, both consisting of the same routine and taking place in a 5m x 6m draft free room, with an ambient temperature of 21°C.

## 4.2 Evaluation Results

The prototype was evaluated according to a number of criteria that follow directly from user requirements. The criteria were: ease of use, data yield, accuracy, robustness, communication range, and information gain.

*Ease of Use.* Instrumentation systems, particularly those used for bomb disposal missions, are expected to have stringent ease of use requirements as they should be transparent to the user and should not interfere with the mission. Ease of use was assessed here subjectively by comparing the ease of application of the sensor packages with sensor mountings for a wired data logger.

As mentioned previously some of the sensor packages have been integrated into clothing, whilst some (neck, thigh, and calf) have been taped to the skin. It is expected that clothing integrated sensors will be less accurate than ones taped to the skin because contact with the skin surface will change when the subject is moving. While avoiding the problem of inconsistent contact, taping on sensors, on the other hand, means that they are less convenient to apply and remove. In comparison to using a standard wired data logger, the wireless system mounting takes considerably less time and has been found to be more comfortable by experimental subjects. Further revisions of the prototype will have all sensor packages mounted on individual elasticated straps, ensuring both firm contact and comfort.

*Data Yield* is a measure of the proportion of data captured. Wireless sensing systems are inherently prone to low yields due to both transmission errors and sensor faults. For the system here, during experimentation, no packets were lost in transmission, however 5% of the sensor samples were found to be out of range (95% yield). Most of the out of range values were from particular sensor packages (3.3% from the worst two), with several sensor packages having no out of range values at all. It is likely that the erroneous values were introduced by I2C bus transmission errors. In comparison, there were no errors apparent in the wired data logger values apart from the chest sensor, which produced incorrect values 61% of the time (39% yield for this sensor, 88% over all data logger sensors).

*Accuracy* is a measure of how closely the sensor data obtained corresponds to the underlying physical phenomena being sensed. As the data logger results in figure 9 show, calibration is needed. Discretisation due to the 12 bit resolution causes some information loss that is offset by sampling frequently. The system is currently being calibrated against a newer data logger instrument.

*Robustness* is particularly important for this system as the intended usage scenario involves it functioning in an environment where it may be subjected to large mechanical shocks and radio frequency interference (RFI). The activity regime here reproduces shocks roughly equivalent to normal application usage and the prototype functioned correctly throughout all trials. As yet, no RFI testing has been carried out.

*Communication Range* is a measure of how far the subject can roam from the monitoring station

without losing communications. In line-of-sight tests, a range of 50 metres was achieved, whilst non-line-of-sight range (through several walls) was about 10 metres. Bluetooth communication will be replaced by ZigBee in the next revision of the prototype.

*Information Gain* is a measure of the benefit of the system in terms of providing more (or better) information about the subject. The prototype has demonstrated two advantages. First, by being untethered, it allows data gathering to occur in the field. Second, it provides a means for real-time remote monitoring and actuation as opposed to offline data acquisition, which is the only role fulfilled by the wired data logger.

### 4.3 Data Analysis

A summary of temperature data obtained from all sensors for a sample run is given in the series of graphs in figure 9. Data was recorded during the experimentation using both the prototype system and a wired-in data logger. These graphs show that the two systems agree in terms of the changes in temperature. It is important to note that the shapes of the graphs, not the exact temperatures measured, are compared here as the sensors on the prototype system had not yet been fully calibrated. The graphs show sensed temperatures over a period of time starting from the third activity (crawling) through to the fifth activity (sitting) followed by a repeat from the first (walking on a treadmill) through to the last again (see section 4.1). In the graphs, the start and end times for each activity are indicated by a vertical bar and the activity number (starting with 3) is given in between each set of bars. Note that there were some rest periods between activities, which are left unmarked.

The aim of the experimentation carried out was two-fold. First, the system under development was compared, in terms of the criteria discussed previously, with a commercially available, wired-in data logger. Second, the data obtained from the two systems was compared to check for consistency. The positioning of the sensors, with several locations having more than one sensor, also meant that the data from the system could be compared internally.

While detailed interpretation of the physiological meaning of the data obtained is beyond the scope of this paper, the data gathered is meaningful in the context of the developed application as follows: 1) Large variations in the skin temperature on some of the sites monitored (maximum three degrees C over 30 minutes) indicate the need for both monitoring and accurate cooling actuation; 2) There are uncorrelated skin temperature variations over the sites monitored

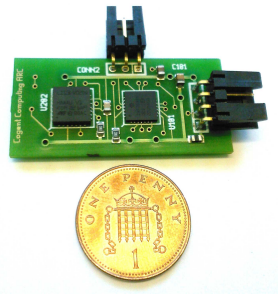


Figure 8: Enhanced sensor package, which is comprised of a PIC processor, 3-DOF accelerometer, I2C buffer, and temperature sensor.

stressing the need for distributed and detailed measurement (as opposed to single point measurement performed by most developed BSN systems); 3) From the graphs, the relationship between activity and skin temperature at different sites is not an obvious one. (An example here is the sudden dip in temperature which occurs for all chest sensors during crawling (activity 3). Crawling is strenuous with the suit on, so this result is surprising.) This indicates the need for added sensing such as humidity and posture information in order to predict the physiological effects of wearing the suit during such exercise regimes. The next prototype, currently under production contains such enhanced sensor packages.

## 5 CONCLUSIONS AND FUTURE WORK

WSN technology is clearly an enabler for detailed measurement in domains such as the one discussed in this paper, domains which are currently not sufficiently understood and lack the necessary instrumentation to further scientific investigation.

Experimental results obtained with a detailed, WSN-based temperature monitoring instrument showed that 1) under a set of activities typical to a bomb disposal mission, skin temperatures for different parts of the body (arms, thigh, chest, and so forth) vary differently thus there is value in sampling at many points; 2) skin temperatures exhibit large variations leading potentially to heat exhaustion hence the need for health monitoring of subjects; 3) autonomous feedback control of the in-suit cooling system, based on a detailed map of how temperature is changing over time, is enabled by the prototype developed so far but more work is needed to determine how best to respond to changes in temperature to ensure that the wearer is kept comfortable.

In the next revision of the prototype, it is planned

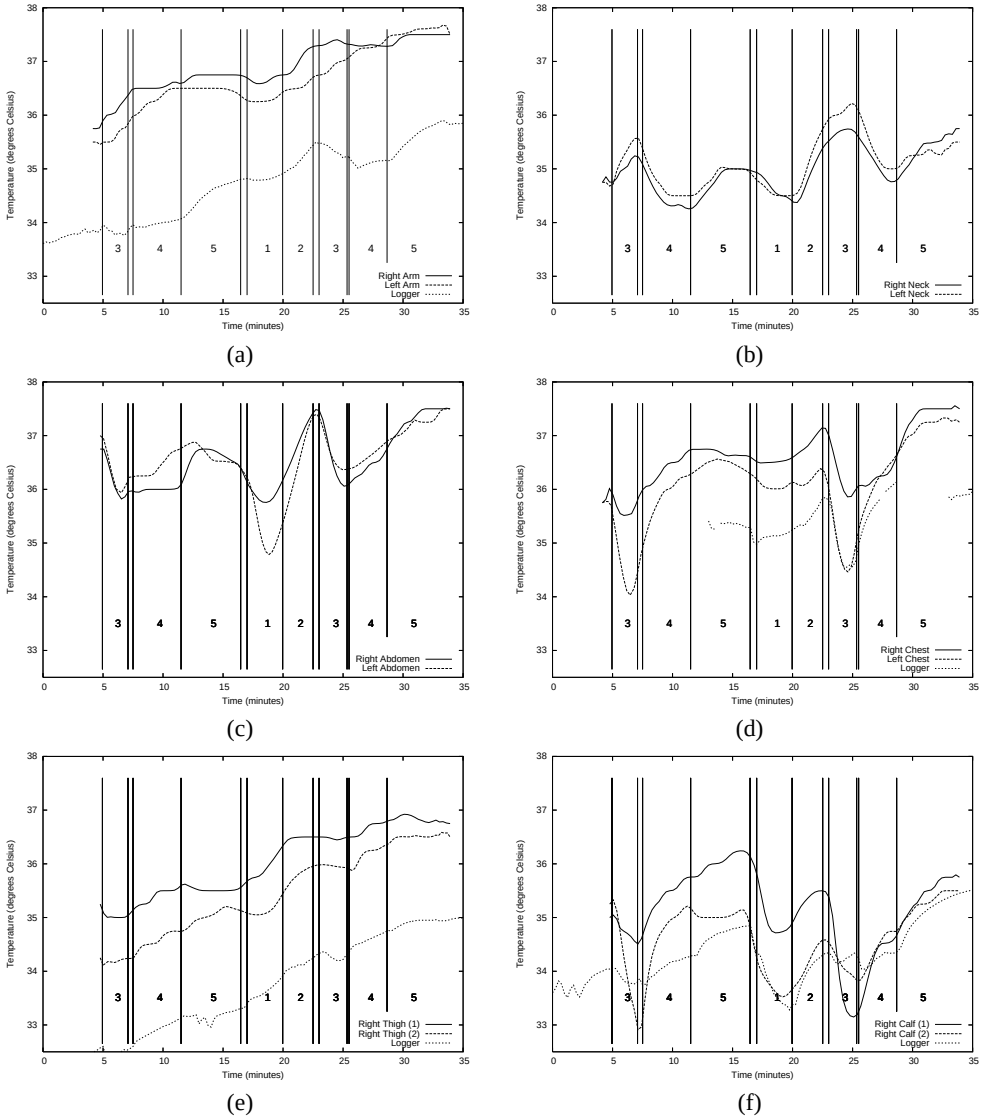



Figure 9: Skin temperature over time for (a) arm, (b) neck, (c) abdomen, (d) chest, (e) thigh, and (f) calf sites. The two leg sensors (thigh and calf positions) were placed on the right leg only. For several skin sites, temperature values were also obtained using a wired-in data logger (denoted "Logger"). The vertical lines in each graph show the start and end of activities. Each activity is represented by a number.

to integrate temperature sensors into a multi-modal sensor board designed in-house, as shown in figure 8. Each board has a temperature sensor and an accelerometer, along with a PIC micro-controller. The two sensors allow the combined monitoring of temperature and acceleration data at any point on a subject's body. The acceleration data will be used for posture identification in further work, which will allow enhanced, activity based remote visualisation of the subject and lead to improved estimates about how the thermal state and thus comfort of the subject is changing. This will hence improve the timeliness and appropriateness of autonomous cooling decisions.

## ACKNOWLEDGEMENTS

The authors wish to thank Doug Thake and his students for the use of and assistance with the Coventry University Health and Life Sciences laboratory and equipment. The authors also wish to thank Bob Newman for his contribution to hardware development in the early stages of this project.

## REFERENCES

Gao, T., Greenspan, D., Welsh, M., Juang, R. R., and Alm, A. (2005). Vital signs monitoring and patient tracking over a wireless network. In *27th Annual International Conference of the IEEE EMBS*, pages 102–105, Shanghai.

Jovanov, E., Raskovic, D., Price, J., Chapman, J., Moore, A., and Krishnamurthy, A. (2001). Patient monitoring using personal area networks of wireless intelligent sensors. *Biomedical Sciences Instrumentation*, 37:373–378.

Keoh, S. L., Dulay, N., Lupu, E., Twidle, K., Schaeffer-Filho, A. E., Sloman, M., Heeps, S., Strowes, S., and Sventek, J. (2007). Self managed cell: A middleware for managing body sensor networks. In *Proceedings of the 4th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services (Mobiquitous)*, Philadelphia, USA.

Lo, B. and Yang, G.-Z. (2005). Architecture for body sensor networks. In *Perspective in Pervasive Computing*, pages 23–28.

Shanks, C. A. (1975). Mean skin temperature during anaesthesia: An assessment of formulae in the supine surgical patient. *British Journal of Anaesthesia*, 47(8):871–876.

Shnayder, V., rong Chen, B., Lorincz, K., Fulford-Jones, T. R. F., and Welsh, M. (2005). Sensor networks for medical care. Technical report, Division of Engineering and Applied Sciences, Harvard University.

Silberstein, E., Bahra, G., and Kattan, J. (1975). Thermographically measured normal skin temperature asymmetry in the human male. *Cancer*, 36(4):1506–1510.

Tapia, E. M., Marmasse, N., Intille, S. S., and Larson, K. (2004). Mites: Wireless portable sensors for studying behavior. In *Proceedings of Extended Abstracts UbiComp 2004: Ubiquitous Computing*.

Thake, C. D. and Price, M. J. (2007). Reducing uncompensable heat stress an a bomb disposal (EOD) suit: a laboratory based assessment. In *Proceedings of the 12th International Conference on Environmental Ergonomics (ICEE 2007)*, Piran, Slovenia. ISBN 978-961-90545-1-2.

Vivometrics (2007). Vivometrics: Better results through non-invasive monitoring. Website. http://www.vivometrics.com; (Online; accessed 16/11/2007).

Walker, W., Polk, T., Hande, A., and Bhatia, D. (2006). Remote blood pressure monitoring using a wireless sensor network. In *6th Annual IEEE Emerging Information Technology Conference*, Dallas, Texas.

# A GUARANTEED STATE BOUNDING ESTIMATION FOR UNCERTAIN NON LINEAR CONTINUOUS TIME SYSTEMS USING HYBRID AUTOMATA

Nacim Meslem, Nacim Ramdani and Yves Candau
*CERTES EA 3481 Université Paris 12-Val de Marne, 61 av. G. de Gaulle, Créteil, France*
*INRIA Sophia Antipolis - Méditerranée - Antenne de Montpellier (LIRMM)*
*161 rue Ada - 34392 Montpellier cedex, France*
*meslem@univ-paris12.fr, nacim.ramdani@inria.fr, candau@univ-paris12.fr*

Keywords: State estimation, non linear systems, bounded error context, non linear differential equations, interval analysis, guaranteed numerical integration, hybrid automata.

Abstract: This work is about state estimation in the bounded error context for non linear continuous time systems. The main idea is to seek to estimate not an optimal value for the unknown state vector but the set of feasible values, thus to characterize simultaneously the value of the vector and its uncertainty. Our contribution resides in the use of comparison theorems for differential inequalities and the analysis of the monotonicity of the dynamical systems with respect to the uncertain variables. The uncertain dynamical system is then bracketted between two hybrid dynamical systems. We show how to obtain this systems and to use them for state estimation with a prediction-correction type observer. An example is given with bioreactors.

## 1 INTRODUCTION

State estimation with continuous dynamic systems is recognized as problem of great importance in practice. Indeed, to apply advanced methods for the control or the diagnosis of dynamical systems one often needs to compute on-line their internal state. Generally the direct measurement of this state by means of sensors may not be available for various reasons such as physical, practical, economic, ... etc. However, it is possible to carry out this task by software sensors, i.e. observers or estimator which can provide on line an estimate of the real state system, under certain conditions of observability (Hermann and J.Krener, 1977; Hermann, 1963).

In fact, there are always uncertainties in the mathematical models used for characterizing the system under study. Consequently, the classical approaches for building observers are insufficient (Dochain, 2003). Thus, a new approach was developed recently in a deterministic set-membership context, which aims to reconstruct all the state trajectories which are consistent with both the uncertain models and the uncertain measurements.

This approach can be used easily when measurements are available at discrete time. It is of *prediction-correction* type: (*i*) The prediction phase consists in computing a guaranteed over (conservative) approximation of the reachable state space generated by the uncertain system, (*ii*) and the correction phase consists in removing from this over approximation all the part which are not consistent with feasible measurements domains, each time a measurement is available.

In the literature, several geometrical forms are used to implement this set-membership approach with linear systems. For example, parallelotopes (Chisci et al., 1996), ellipsoidal (Chernousko, 2005) and zonotopes (Combastel, 2005).

With nonlinear systems, guaranteed numerical integration method for the ordinary differential equation (ODE) (Nedialkov, 1999) based on intervals Taylor models (Moore, 1966) was used recently to solve this estimation problem (Raïssi et al., 2004). Generally, the wrapping effect (Moore, 1966; Nedialkov, 1999) associated with the intervals representation of uncertain variables limits considerably their use in order to deal with practical cases. Thereafter, in order to circumvent the wrapping effect, the authors of (Kieffer and Walter, 2006) used the Müller's existence theorem (Müller, 1926; Walter, 1997) as a tool for deriving a guaranteed enclosure for an uncertain dynami-

cal system between two deterministic dynamical systems. The main difficulty resides in the definition of the bracketing systems. Our contribution is in the continuation of these work since we show how to build deterministic hybrid dynamical systems as bracketing systems and thus make it possible to use the Müller theorem with a larger class of nonlinear dynamical systems.

This work is organized as flow. In the second section we present the context and the main ideas of set-membership estimation. Then, we will recall the Müller's theorem in section 3. We introduce the hybrid bracketing approach for uncertain dynamical systems in section 4. Finally we illustrate our method on a model drawn from bioreactors domain in section 5.

# 2 PROBLEM STATEMENT

## 2.1 Context

Let us consider the uncertain continuous dynamic system (1) where uncertainties are naturally represented by bounded intervals with *a priori* known bounds,

$$\left\{ \begin{array}{c} \dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x},\mathbf{u},[\mathbf{p}],t) \\ \mathbf{y}(t) = \mathbf{g}(\mathbf{x},\mathbf{u},[\mathbf{p}],t) \\ \mathbf{x}(t_0) \in [\mathbf{x}_0] \subset \mathbb{D} \end{array} \right\} \quad (1)$$

where $t \in [t_0, T]$, $\mathbf{f} \in C^{k-1}(\mathbb{D} \times \mathbb{U} \times [\mathbf{p}])$, $\mathbb{D} \times \mathbb{U} \times [\mathbf{p}] \subseteq \mathbb{R}^{n+n_u+n_p}$ is an open set; $n$, $n_u$, $m$ and $n_p$ are the dimension of respectively the state vector $\mathbf{x}$, the input vector $\mathbf{u}$, the output vector $\mathbf{y}$ and the parameter vector $\mathbf{p}$. The functions $\mathbf{f} : \mathbb{D} \times \mathbb{U} \times [\mathbf{p}] \rightarrow \mathbb{R}^n$ and $\mathbf{h} : \mathbb{D} \times \mathbb{U} \times [\mathbf{p}] \rightarrow \mathbb{R}^m$ are possibly nonlinear. The initial state $\mathbf{x}_0$ is assumed to belong to a prior known set $[\mathbf{x}]$. We assume that measurements $\mathbf{y}_j$ of the output vector are available at sampling times $t_i \in \{t_1, t_2, \ldots, t_n\}$ in $I = [t_0, t_{n_T}]$. Note that the sampling interval needs not be constant. The measurement noise is a discrete time signal assumed additive and bounded with known bounds. Denote $\mathbb{E}_j$ a feasible domain for output error at time $t_j$: the feasible domain for model output at time $t_j$ is then given by

$$\mathbb{Y}_j = \mathbf{y}_j + \mathbb{E}_j \quad (2)$$

Under these considerations, estimating the state vector $\mathbf{x}$ consists in determining an upper approximation of the set $\mathbb{X}(t)$ of all acceptable state trajectories

$$\mathbb{X}(t) = \left\{ \begin{array}{c} \mathbf{x}(t) \mid (\forall t \in I \; \dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x},\mathbf{u},[\mathbf{p}],t)) \\ \wedge \\ (\forall t_j \in \{t_1, t_2, \ldots, t_{nT}\}, \\ \mathbf{x}(t_j) \in (\mathbf{g}^{-1}(\mathbb{Y}(t_j),\mathbf{u},[\mathbf{p}]) \cap [\mathbf{x}_j])) \end{array} \right\} \quad (3)$$

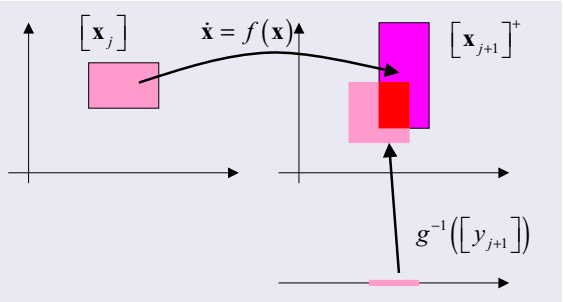



Figure 1: Prediction and correction phases.

## 2.2 Principle: Prediction-correction Method

Figure 1 shows the principle of the prediction and correction phases for this set-membership approach, between two successive time measurement indexes $t_j$ and $t_{j+1}$. Indeed, by using one of the guaranteed numerical simulation methods for uncertain ODEs, the *prediction* phase computes an guaranteed over enclosure $[\mathbf{x}_{j+1}]^p$ for all solutions of (1) at the time $t_{j+1}$ with $(t_j, [\mathbf{x}_j])$ as initial conditions.

$$\mathbf{x}(t_{j+1}; t_j, [\mathbf{x}_j]) \subseteq [\mathbf{x}_{j+1}]^p. \quad (4)$$

The *correction* phase uses set inversion, consistency techniques and intervals analysis (Jaulin et al., 2001) to characterize the reciprocal image $[\mathbf{x}_{j+1}]^{inv}$ at time $t_{j+1}$ of the admissible measurement domain $\mathbb{Y}_{j+1}$ by the output model $\mathbf{g}$

$$[\mathbf{x}_{j+1}]^{inv} = \mathbf{g}^{-1}([\mathbf{y}_{j+1}], [\mathbf{p}]) \quad (5)$$

where $[\mathbf{y}_{j+1}] = \mathbb{Y}_{j+1}$.

Thereafter, it contracts the predicted state intervals vector by comparing it with the reciprocal image $[\mathbf{x}_{j+1}]^{inv}$ and eliminating the inconsistent state vectors

$$[\mathbf{x}_{j+1}]^c = [\mathbf{x}_{j+1}]^{inv} \cap [\mathbf{x}_{j+1}]^p \quad (6)$$

Thus for the next measurement the predication phase will be initialized by $[\mathbf{x}_{j+1}] = [\mathbf{x}_{j+1}]^c$. In fact, by repeating these two phases each time a new measurement is available, one improves considerably precision of the guaranteed over approximation of the set $\mathbb{X}(t)$.

The algorithm below shows the process for this set-membership estimation approach

**Algorithm : Prediction_Correction_estimation**

1. **Input**: $([\mathbf{x}_0],[\mathbf{p}],\mathbf{f},\mathbf{g},[\mathbf{y}_1],\ldots,[\mathbf{y}_{nT}])$
2. $t_j = t_0$; $[\mathbf{x}_j] = [\mathbf{x}_0]$;
3. **while** $(t_j < t_{nT})$ **do**
4. $\{t_{j+1},[\mathbf{x}_{j+1}]^p\}$=**Validated_Integration**$([\mathbf{x}_j],[\mathbf{p}],t_j)$;
5. $[\mathbf{x}_{j+1}]^{inv} = \mathbf{g}^{-1}([\mathbf{y}_{j+1}],[\mathbf{p}])$;
6. $[\mathbf{x}_{j+1}]^c = [\mathbf{x}_{j+1}]^{inv} \cap [\mathbf{x}_{j+1}]^p$;
7. $[\mathbf{x}_{j+1}] = [\mathbf{x}_{j+1}]^c$
8. $j = j+1$;
9. **end**
10. **Output**: $\mathbb{X}(t)$.

# 3 MÜLLER'S THEOREM

In this section, we introduce an approach for bracketing an uncertain dynamical systems when both the initial state and parameter vectors are defined by boxes (intervals vector). The main idea consists in building a lower and an upper dynamical system which involve no uncertainty and enclose in a guaranteed way, the all state trajectories generated by the original uncertain system. This approach relies on comparison theorems for differential inequalities (Smith, 1995; Hirsch and Smith, 2005), and in particular the work of Müller (Müller, 1926; Marcelli and Rubbioni, 1997).

*Theorem*:(Müller, 1926; Kieffer and Walter, 2006)
Consider the dynamical system

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x},\mathbf{p},\mathbf{u}(t)), \tag{7}$$

where function $\mathbf{f}$ is continuous over a domain $\mathbb{T}$ defined by

$$\mathbb{T}: \begin{cases} \omega(t) \le \mathbf{x}(t) \le \Omega(t) \\ \underline{\mathbf{p}} \le \mathbf{p} \le \overline{\mathbf{p}} \\ t_0 \le t \le t_{n_T} \end{cases} \tag{8}$$

Functions $\omega_i(t)$ and $\Omega_i(t)$ are continuous over $[t_0,t_{n_T}]$ for all $i$ and satisfy the following properties

1. $\omega(t_0) = \underline{\mathbf{x}}_0$ and $\Omega(t_0) = \overline{\mathbf{x}}_0$
2. the left derivatives $D^-\omega_i(t)$ and $D^-\Omega_i(t)$ and the right derivatives $D^+\omega_i(t)$ and $D^+\Omega_i(t)$ of $\omega_i(t)$ and $\Omega_i(t)$ are such that

$$\forall i,\ D^{\pm}\omega_i(t) \le \min_{\mathbb{T}(t)} f_i(\mathbf{x},\mathbf{p},t) \tag{9}$$

$$\forall i,\ D^{\pm}\Omega_i(t) \ge \max_{\mathbb{T}(t)} f_i(\mathbf{x},\mathbf{p},t) \tag{10}$$

where $\underline{\mathbb{T}}(t)$ is the subset of $\mathbb{T}(t)$ defined by

$$\underline{\mathbb{T}}_i: \begin{cases} x_i = \omega_i(t) \\ \omega_j(t) \le x_j \le \Omega_j(t),\ j \ne i \\ \underline{\mathbf{p}} \le \mathbf{p} \le \overline{\mathbf{p}} \end{cases} \tag{11}$$

and where $\overline{\mathbb{T}}(t)$ is the subset of $\mathbb{T}(t)$ defined by

$$\overline{\mathbb{T}}_i: \begin{cases} x_i = \Omega_i(t) \\ \omega_j(t) \le x_j \le \Omega_j(t),\ j \ne i \\ \underline{\mathbf{p}} \le \mathbf{p} \le \overline{\mathbf{p}} \end{cases} \tag{12}$$

Then for all $\mathbf{x}_0 \in [\underline{\mathbf{x}}_0,\overline{\mathbf{x}}_0]$, $\mathbf{p} \in [\underline{\mathbf{p}},\overline{\mathbf{p}}]$, system (1) admits a solution $\mathbf{x}(t)$ that stays in the domain

$$\mathbb{X}: \begin{cases} t_0 \le t \le t_{n_T} \\ \omega(t) \le \mathbf{x}(t) \le \Omega(t) \end{cases} \tag{13}$$

and takes the value $\mathbf{x}_0$ at $t_0$. If, in addition, for all $\mathbf{p} \in [\underline{\mathbf{p}}_0,\overline{\mathbf{p}}_0]$, function $\mathbf{f}(\mathbf{x},\mathbf{p},t)$ is Lipschitzian with respect to $\mathbf{x}$ over $\mathbb{D}$ then this solution is unique for any given $\mathbf{p}$.

Finally, an enclosure for the solution of (7) is given by

$$\forall t \in [t_0,t_{n_T}],\quad [\mathbf{x}](t) = [\omega(t),\Omega(t)] \tag{14}$$

The main difficulty is to obtain suitable bracketing functions $\omega(t)$ and $\Omega(t)$ in the general case. However, when the components of $\mathbf{f}$ are monotonic with respect to each parameter and each state vector component, it is quite easy to define these systems (Kieffer et al., 2006), while avoiding possible divergence that may occur when both upper and lower components of the parameter/state vector appear simultaneously in the same expression of the components of the bracketing systems (Ramdani et al., 2006).

*Rule 1 - Use of monotonicity property* (Kieffer et al., 2006)
In order to build the upper system, i.e. the one which yields the upper solution $\Omega(t)$, one can replace in the formal expression of $f_i$, $x_i$ by $\Omega_i$, $x_j$ $(j \ne i)$ by $\Omega_j$ if $\frac{\partial f_i}{\partial x_j} \ge 0$ or by $\omega_j$ if $\frac{\partial f_i}{\partial x_j} \le 0$ and $p_k$ by $\overline{p}_k$ if $\frac{\partial f_i}{\partial p_r} \ge 0$ or by $\underline{p}_k$ if $\frac{\partial f_i}{\partial p_k} \le 0$. The components of the lower system, i.e. the one which yields the lower solution $\omega(t)$ are derived by reversing monotonicity conditions.

Obviously $\omega(t)$ and $\Omega(t)$ are in general, solutions of a system of coupled differential equations, i.e.

$$\begin{cases} \dot{\omega}(t) = \underline{\mathbf{f}}(\omega,\Omega,\underline{\mathbf{p}},\overline{\mathbf{p}},t) \\ \dot{\Omega}(t) = \overline{\mathbf{f}}(\omega,\Omega,\underline{\mathbf{p}},\overline{\mathbf{p}},t) \\ \omega(t_0) = \underline{\mathbf{x}}_0 \\ \Omega(t_0) = \overline{\mathbf{x}}_0 \end{cases} \tag{15}$$

which involves no *uncertain* quantity. Therefore interval Taylor models such as the one presented in (Nedialkov, 1999) can be used for efficiently solving (15). Indeed when these methods are used for solving differential equations with no uncertainty, they are usually able to curb the pessimism induced by the wrapping effect, even over long integration time.

# 4 HYBRID BRACKETING SYSTEM

Now, one address the case of uncertain dynamical systems (1), for which the signs of the partial derivatives $\partial f_i/\partial p_k$ and $\partial f_i/\partial x_j$ change along the integration time interval $[t_0, t_{n_T}]$. In such a case, the uncertain system (1) admits an enclosure over each time interval where functions $f_i$ are monotonic with respect to variables $p_k$ and $x_j$. Therefore both upper and lower bounding systems are defined by piecewise nonlinear ODEs and can thus be regarded as hybrid dynamical systems. Thus, they can be modeled by an *hybrid automaton*(Alur et al., 1995).

So to compute a guaranteed enclosure of reachable state space generated by the uncertain system (1), we will built hybrid system of $l$ continuous dynamic modes which satisfied locally conditions imposed by *rule 1*

$$\mathbb{M} = \{M_1, M_2, ..., M_l\} \tag{16}$$

and which given in state space representation by

$$\begin{cases} \dot{\omega}(t) = \underline{\mathbf{f}}_{M_i}(\omega, \Omega, \underline{\mathbf{p}}, \overline{\mathbf{p}}, t) \\ \dot{\Omega}(t) = \overline{\mathbf{f}}_{M_i}(\omega, \Omega, \underline{\mathbf{p}}, \overline{\mathbf{p}}, t) \end{cases} \text{for } i = 1, ..., l \tag{17}$$

Hence, the evolution of this hybrid system is controlled by the sign changes of the partial derivatives $\partial f_i/\partial p_k$ and $\partial f_i/\partial x_j$ which represents the guard conditions which authorize the transitions between the continuous bracketing modes. Thus, for a given initial conditions the execution of this hybrid automata makes it possible to obtain a guaranteed upper approximation of the reachable state space of the continuous time system (1).

**Example**:

Let us consider the following system

$$\dot{x}(t) = f(x, [p]). \tag{18}$$

According to the sign of $\partial f/\partial p$, the system (18) admitted two possible bracketing modes

$$if \;\; \partial f/\partial p \leq 0$$

$$then \;\; M_1 = \begin{cases} \dot{\Omega}(t) = f_{M_1}(\Omega, \underline{p}) \\ \dot{\omega}(t) = f_{M_1}(\omega, \overline{p}) \end{cases} \tag{19}$$

$$else \;\; M_2 = \begin{cases} \dot{\Omega}(t) = f_{M_2}(\Omega, \overline{p}) \\ \dot{\omega}(t) = f_{M_2}(\omega, \underline{p}). \end{cases} \tag{20}$$

and its hybrid bracketing automata is represented in the figure 2

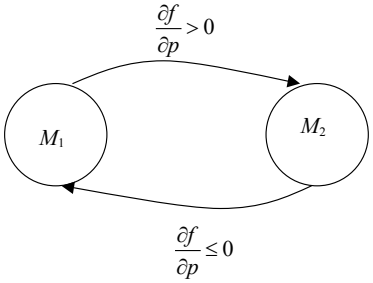



Figure 2: Hybrid automata.

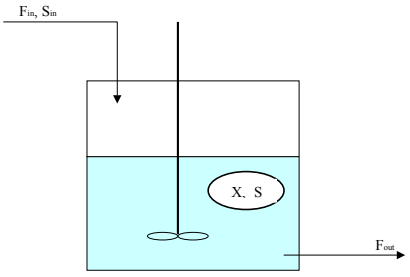



Figure 3: General representation of a bioreactor.

# 5 APPLICATION

## 5.1 Model

Generally, the mathematical model of biotechnological processes is difficult to establish with accuracy, that is due to the living behavior of the bacteria represented by a complex poorly known function of the bioreactor state. In this section we consider a simple model where only one population of bacteria is taken into account. In this context, to describe the state of the bioreactor, two state variables are necessary, the first one represents the bacteria concentration called biomass and denoted $X$, the second one represents the substrate concentration, denoted $S$. Thus the model below shows the evolution of the biomass by consuming the polluting body $S$

$$\begin{cases} X(t) = \mu(S)X - \alpha DX \\ S(t) = -k_1\mu(S)X + D(S_{in} - S) \end{cases} \tag{21}$$

where $\mu(s)$ is the growth rate of biomass modeled by the Haldane law:

$$\mu(S) = \mu_0 \frac{S}{S + k_s + S^2/k_i} \tag{22}$$

with the uncertain bounded parameter $\mu_0$

$$\underline{\mu}_0 \leq \mu_0 \leq \overline{\mu}_0.$$

This bioreactor is fed by a solution containing substrate in concentration $S_{in}$ which is not exactly measured

$$\underline{S}_{in}(t) \leq S_{in}(t) \leq \overline{S}_{in}(t)$$

and we suppose that the biomass is accessible to measurement

$$y(t) = X(t).$$

## 5.2 Building Hybrid Bracketing System

For this application, according to the sign of the derivative of $\mu$ with respect to $S$,

$$sign(\frac{d\mu(S)}{dS}) \begin{cases} > 0 \text{ if } S < \sqrt{k_s k_i},\ \forall S \in [S] \\ \leq 0 \text{ if } S \geq \sqrt{k_s k_i},\ \forall S \in [S] \\ \text{ambiguous if } \sqrt{k_s k_i} \in [S] \end{cases} \quad (23)$$

system (21) allows three possible modes for the bracketing. The first mode, $M_1 = 1$, corresponds to the intervals time when this derivative is negative

$$(M_1 = 1) \begin{cases} \overline{X}(t) = \overline{\mu}(\overline{S})\overline{X} - \alpha D\overline{X} \\ \overline{S}(t) = -k_1\underline{\mu}(\overline{S})\underline{X} + D(\overline{S}_{in} - \overline{S}) \\ \underline{X}(t) = \underline{\mu}(\underline{S})\underline{X} - \alpha D\underline{X} \\ \underline{S}(t) = -k_1\overline{\mu}(\underline{S})\overline{X} + D(\underline{S}_{in} - \underline{S}) \end{cases} \quad (24)$$

and the second mode, $M_2 = 2$, is linked to intervals time where this derivative is positive

$$(M_2 = 2) \begin{cases} \overline{X}(t) = \overline{\mu}(\underline{S})\overline{X} - \alpha D\overline{X} \\ \overline{S}(t) = -k_1\underline{\mu}(\overline{S})\underline{X} + D(\overline{S}_{in} - \overline{S}) \\ \underline{X}(t) = \underline{\mu}(\overline{S})\underline{X} - \alpha D\underline{X} \\ \underline{S}(t) = -k_1\overline{\mu}(\underline{S})\overline{X} + D(\underline{S}_{in} - \underline{S}). \end{cases} \quad (25)$$

Finally, the third mode $M_3 = 0$ is associated to case when the sign of this derivative is ambiguous. In this case either one uses guaranteed integration methods based on interval Taylor models to bracket (21), or if this is possible, one finds a trivial bracketing for $\mu(s)$. For example,

$$\underline{\mu}_0 \frac{\underline{S}}{\underline{S} + k_s + \underline{S}\overline{S}/k_i} \leq \mu(S) \leq \overline{\mu}_0 \frac{\overline{S}}{\overline{S} + k_s + \underline{S}\overline{S}/k_i}.$$

Hence, one propose the below differential equations system for the third mode $M_3 = 0$

$$(M_3 = 0) \begin{cases} \overline{X}(t) = \overline{\mu}_0 \frac{\overline{S}}{\overline{S} + k_s + \underline{S}\overline{S}/k_i}\overline{X} - \alpha D\overline{X} \\ \overline{S}(t) = -k_1\underline{\mu}(\overline{S})\underline{X} + D(\overline{S}_{in} - \overline{S}) \\ \underline{X}(t) = \underline{\mu}_0 \frac{\underline{S}}{\underline{S} + k_s + \underline{S}\overline{S}/k_i}\underline{X} - \alpha D\underline{X} \\ \underline{S}(t) = -k_1\overline{\mu}(\underline{S})\overline{X} + D(\underline{S}_{in} - \underline{S}). \end{cases} \quad (26)$$

Now, function **Validated_Integration** in algorithm **Prediction_Correction_Estimation** selects on line the local bracketing mode according to the sign of $\frac{d\mu(.)}{dS}$ and then uses guaranteed numerical integration methods for ODE based on intervals Taylor models for solving (24), (25) or (26).

## 5.3 Results of Simulation

The data considered in this example are as follows: $\alpha = 0.5$, $k = 42.14$, $k_s = 9.28 mmol/l$, $k_i = 256 mmol/l$, $\mu_0 \in [0.64, 0.84]$, $X_0 \in [0>, 10]$, $S_0 \in [0, 100]$, $S_{in}(t) \in ([62, 68] + 15\cos(1/5t))$,

$$D(t) = \begin{cases} 2 \text{ si } 0 \leq t \leq 5 \\ 0.5 \text{ si } 5 \leq t \leq 10 \\ 1.14 \text{ si } 10 \leq t \leq 20, \end{cases}$$

and the feasible measurement domain

$$\mathbb{Y}(t_j) = [0.98 y_m(t_j), 1.02 y_m(t_j)]$$

with a constant measurements time step $t_{j+1} - t_j = 2$.

The red continuous lines curves in figures 4 and 5 show the guaranteed enclosure of all the possible state trajectories of (21) which are compatible with the model and its uncertainties and the acceptable domain of the discrete measurements signal. The discontinuous blue curves represent the real state of (21) which corresponds to the following values of the uncertain parameters and the initial state: $\mu_0 = 0.74$, $X_0 = 5$, $S_0 = 40$ and $S_{in}(t) = 65 + 15\cos(1/5t)$.

So figure 6 shows the actual commutations between the three bracketing modes as obtained during the simulation period. This represents the evolution of the discrete component of the hybrid automata used to bracket the state flow generated by the uncertain system (21).

As a conclusion, guaranteed numerical integration methods for ODE based on the interval Taylor models fail to give non divergent enclosures after few integration step because of the large magnitude in uncertainty in both parameter vector and initial state vector. In addition, *rule 1* is not applicable over all the simulation period because the sign of the partial derivative $\frac{\partial \dot{X}}{\partial S}$ changes with time. Hence, our hybrid bracketing method is an important alternative to solve set membership estimation problems of this kinds.

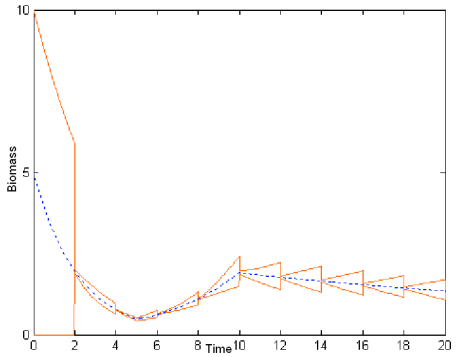



Figure 4: The both real and estimated evolution the biomass.

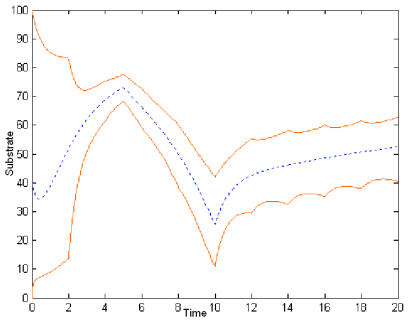



Figure 5: The both real and estimated evolution the substrate.

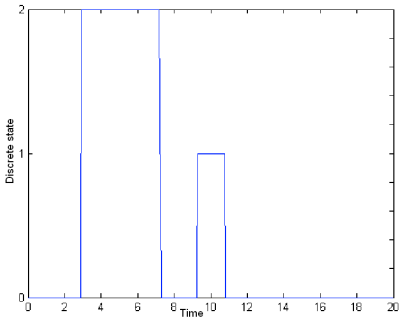



Figure 6: Indices of the three bracketing modes used over the simulation period.

# 6 CONCLUSIONS

In this communication, we wanted to show that by using the Müller's theorem and by analyzing the monotonicity of the uncertain dynamical system with respect to both the uncertain variables and parameters, one is able to solve the state membership estimation problem for a large class of uncertain dynamical systems. Indeed, the method presented makes it possible to circumvent the propagation of the pessimism due to the wrapping effect which generally causes the divergence of guaranteed numerical integration methods based on interval Taylor models when used with uncertain ODEs. In the future, we wish to extend this approach for systems with higher dimension and also to hybrid dynamical systems.

# REFERENCES

Alur, R., Courcoubetis, C., Halbwachs, N., Henzinger, T., Ho, P.-H., Nicollin, X., Olivero, A., Sifakis, J., and Yovine, S. (1995). The algorithmic analysis of hybrid systems. *Theoretical Computer Science*, 138:3–34.

Chernousko, F. (2005). Ellipsoidal state estimation for dynamical systems. *Nonlinear analysis*.

Chisci, L., Garulli, A., and Zappa, G. (1996). Recursive state bounding by parallelotopes. *Automatica*, 32:1049–1055.

Combastel, C. (2005). A state bounding observer for uncertain nonlinear continuous-time systems based on zonotopes. In *44th IEEE Conference on decision and control and European control conference ECC 2005*.

Dochain, D. (2003). State and parameter estimation in chemical and biochemical processes: a tutorial. *Journal of process control*, 13:801–818.

Hermann, R. (1963). On the accessibility problem in control theory. In *in Int. Symp. on nonlinear differential equations and nonliear mechanics*, pages 325–332, New York. Academic Press.

Hermann, R. and J.Krener, A. (1977). Nonlinear controllability and observability. *Transactions on Automatic Control*, AC-22:728–740.

Hirsch, M. and Smith, H. (2005). Monotone dynamical systems. In Canada, A., Drabek, P., and Fonda, A., editors, *Handbook of Differential Equations, Ordinary Differential Equations*, volume 2, chapter 4. Elsevier.

Jaulin, L., Kieffer, M., Didrit, O., and Walter, E. (2001). *Applied interval analysis: with examples in parameter and state estimation, robust control and robotics*. Springer-Verlag, London.

Kieffer, M. and Walter, E. (2006). Guaranteed nonlinear state estimation for continuous-time dynamical models from discrete-time measurements. In *Proceedings 6th IFAC Symposium on Robust Control*, Toulouse.

Kieffer, M., Walter, E., and Simeonov, I. (2006). Guaranteed nonlinear parameter estimation for continuous-time dynamical models. In *Proceedings 14th IFAC Symposium on System Identification*, pages 843–848, Newcastle, Aus.

Marcelli, C. and Rubbioni, P. (1997). A new extension of classical müller theorem. *Nonlinear Analysis, Theory, Methods & Applications*, 28(11):1759–1767.

Moore, R. (1966). *Interval analysis*. Prentice-Hall, Englewood Cliffs.

Müller, M. (1926). Uber das fundamentaltheorem in der theorie der gewöhnlichen differentialgleichungen. *Math. Z.*, 26:619–645.

Nedialkov, N. (1999). *Computing rigorous bounds on the solution of an initial value problem for an ordinary differential equation*. PhD University of Toronto.

Raïssi, T., Ramdani, N., and Candau, Y. (2004). Set membership state and parameter estimation for systems described by nonlinear differential equations. *Automatica*, 40(10):1771–1777.

Ramdani, N., Meslem, N., Raïssi, T., and Candau, Y. (2006). Set-membership identification of continuous-time systems. In *Proceedings 14th IFAC Symposium on System Identification*, pages 446–451, Newcastle, Aus.

Smith, H. (1995). *Monotone dynamical systems: An introduction to the theory of competitive and cooperative systems*. Ams. providence, ri edition.

Walter, W. (1997). Differential inequalities and maximum principles: Theory, new methods and applications. *Nonlinear Analysis, Theory, Methods & Applications*, 30(8):4695–4711.

# RECURSIVE BIAS-COMPENSATING ALGORITHM FOR THE IDENTIFICATION OF DYNAMICAL BILINEAR SYSTEMS IN THE ERRORS-IN-VARIABLES FRAMEWORK

T. Larkowski, J. G. Linden, B. Vinsonneau and K. J. Burnham
*Control Theory and Applications Centre, Coventry University, Prior Street, Coventry, U.K.*
*larkowst@coventry.ac.uk*

Keywords: Bias compensation, Bilinear systems, Errors-in-variables, Recursive estimation, Regularization, System identification.

Abstract: The paper investigates a recursive approach for the bias compensating least squares (BCLS) technique. The method presented is applied to the problem of on-line identification of single-input single-output bilinear models in the errors-in-variables framework. Within this framework the recursive bilinear BCLS algorithm is realized when a bilinear Frisch scheme (BFS) is iteratively applied for the estimation of the parameters of an exemplary bilinear system, giving rise to the exact recursive BFS (ERBFS) method. Moreover, a further extension of the ERBFS incorporating Tikhonov regularization with variable exponential weighting is considered and this is shown to be beneficial in the initial period of the identification procedure.

## 1 INTRODUCTION

The errors-in-variables (EIV) framework addresses the identification of dynamical systems where both the input and the output signals are corrupted by the measurement noise (Söderström, 2007). The EIV approaches are found to be of considerable benefit when the underlying physical laws characterizing the system are of a prime interest, as opposed to the prediction of the external signals (Söderström et al., 2002). In this case, the classical methods based on the least squares (LS) principle such as recursive LS (RLS) or the Kalman filter (Ikonen and Najim, 2002) are shown to yield estimates of the system parameters that are asymptotically biased and, therefore, inconsistent (Zheng, 1998; Söderström, 2007).

In the field of modelling for nonlinear systems, the bilinear system (BS) models have been used to advantage in various practical applications, e.g. control plants, biological and chemical phenomena, earth and sun science, nuclear fission, fault diagnosis and supervision, see (Mohler, 1991; Mohler and Khapalov, 2000) or (Ekman, 2005). Due mainly to the fact that BS models are so widely applicable has prompted the need to extend the EIV approaches developed for linear systems to encompass the BS case.

Recently, a technique for off-line compensation of the bias in the case of dynamical BS, i.e. the bilinear bias compensating LS (BBCLS) scheme has been proposed (Larkowski et al., 2007), upon which a bilinear Frisch scheme (BFS) has been constructed (Larkowski et al., 2008). The focus of this paper is the extension of the BBCLS along with the BFS for the purpose of on-line system identification. The proposed approach consists of a recursively performed update and bias compensation procedure for the data covariance matrices, whilst the BFS equations are applied in an iterative manner at each recursion step. Moreover, a further extension of the BFS incorporating the Tikhonov regularization (TR) technique with a variable exponential weighting is considered. It is shown via simulation studies that use of TR can be of considerable benefit in the initial period of the identification procedure.

The paper is organized as follows: in the second section the mathematical representation of the EIV BS together with the assumptions stated and the notation used are introduced. The third section presents a brief review of the BBCLS and the BFS techniques. In section four a recursive implementation of the BBCLS method is proposed. Subsequently, within the BBCLS framework the BFS technique is applied resulting in the exact recursive BFS (ERBFS) algorithm. The section ends with an extension of the ERBFS that incorporates the TR technique. Section five presents the results of a numerical simulation study involving the proposed algorithms, whilst the overall conclusions and the further work are summarized in section six.

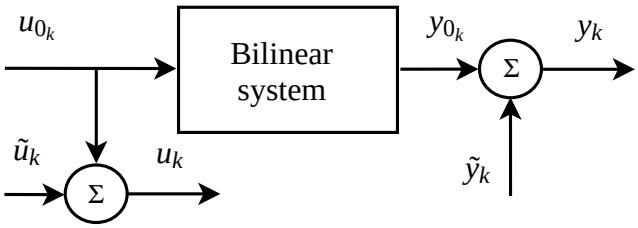



Figure 1: The basic setup for the EIV BS.

## 2 ASSUMPTIONS AND NOTATION

Consider a discrete time-invariant single-input single-output (SISO) class of BS that can be represented by the following input/output difference equation

$$A(q^{-1})y_{0_k} = B(q^{-1})u_{0_k} + \sum_{i=1}^{p}\sum_{j=1}^{r}\eta_{ij}u_{0_{k-i}}y_{0_{k-j}} \qquad (1)$$

with the polynomials $A(q^{-1})$ and $B(q^{-1})$ given by

$$A(q^{-1}) \triangleq 1 + a_1 q^{-1} + \ldots + a_{n_a} q^{-n_a} \qquad (2a)$$

$$B(q^{-1}) \triangleq b_1 q^{-1} + \ldots + b_{n_b} q^{-n_b} \qquad (2b)$$

where $r \leq n_a$, $p \leq n_b \leq n_a$, $q^{-1}$ is the backward shift operator, defined by $x_k q^{-1} \triangleq x_{k-1}$ and $u_{0_k}, y_{0_k}$ are the noise-free input and output sequences, respectively. A diagrammatic illustration of the typical EIV setup for a SISO BS is depicted in Figure 1.

The BS can be classified into three main categories, see (Pearson, 1999) for more details, namely: *a*) subdiagonal $\eta_{ij} = 0 \ \forall j > i$; *b*) diagonal $\eta_{ij} = 0 \ \forall j \neq i$; and *c*) superdiagonal $\eta_{ij} = 0 \ \forall j < i$. Noting that both the subdiagonal and superdiagonal cases include the diagonal case, reference will be solely made here to the diagonal BS (DBS) case for the remainder of the paper. This is due to the fact that DBS exhibit some crucial properties of interest, see (Rao and Gabr, 1984; Liu, 1992) or (Kotta and Nomm, 2003) for a detailed discussion. At the same time DBS are possibly the most commonly utilized class of BS for the purpose of industrial applications, see (Burnham, 1991; Yu, 1996; Martineau et al., 2004).

Without loss of generality, the case when all the diagonal terms in the system (1) are present is considered here with their number given as $n_\eta = p^2$ where $r = p$. The following assumptions are introduced:

**A1.** The DBS is time-invariant, asymptotically stable, strictly stationary, observable and controllable.

**A2.** The system structure, i.e. $n_a$, $n_b$, $p$, is known *a priori*.

**A3.** The true input $u_{0_k} \sim \mathcal{N}(0, \sigma_{u_0})$ is white, persistently exciting and of sufficiently high order.

**A4.** Corrupting input/output noises $\tilde{u}_k \sim \mathcal{N}(0, \sigma_{\tilde{u}})$ and $\tilde{y}_k \sim \mathcal{N}(0, \sigma_{\tilde{y}})$ of unknown variances are additive, white, mutually uncorrelated and uncorrelated with the noise free signals $u_{0_k}$ and $y_{0_k}$, respectively.

Acknowledging A4, it is implied that the measured input and output can be decomposed into their noise free and noise contributions, i.e.

$$u_k \triangleq u_{0_k} + \tilde{u}_k \qquad (3a)$$

$$y_k \triangleq y_{0_k} + \tilde{y}_k \qquad (3b)$$

where $k$ denotes the discrete time index. The identification problem consists of determining the vector

$$\vartheta^T \triangleq \begin{bmatrix}\theta^T & \sigma_{\tilde{u}} & \sigma_{\tilde{y}}\end{bmatrix} \in \mathcal{R}^{n_\theta+2} \qquad (4)$$

where $\theta \in \mathcal{R}^{n_\theta}$ is the parameter vector with $n_\theta = n_a + n_b + n_\eta$ and $\sigma_{\tilde{u}}$, $\sigma_{\tilde{y}}$ are the input and output noise variances, respectively. The parameter vector is defined as:

$$\theta \triangleq \begin{bmatrix}a\\ b\\ \eta\end{bmatrix} \quad a \triangleq \begin{bmatrix}a_1\\ \vdots\\ a_{n_a}\end{bmatrix} \quad b \triangleq \begin{bmatrix}b_1\\ \vdots\\ b_{n_b}\end{bmatrix} \quad \eta \triangleq \begin{bmatrix}\eta_{11}\\ \vdots\\ \eta_{pp}\end{bmatrix} \qquad (5)$$

with $a \in \mathcal{R}^{n_a}$, $b \in \mathcal{R}^{n_b}$, $\eta \in \mathcal{R}^{n_\eta}$ where the extended parameter vector $\bar{\theta}$ is given by

$$\bar{\theta} \triangleq \begin{bmatrix}\bar{a}\\ b\\ \eta\end{bmatrix} \in \mathcal{R}^{n_\theta+1} \qquad \bar{a} \triangleq \begin{bmatrix}1\\ a\end{bmatrix} \in \mathcal{R}^{n_a+1} \qquad (6)$$

The regressor vectors for the measured data, noise-free data and noise are defined, respectively, as:

$$\varphi_k \triangleq \begin{bmatrix}\varphi_{y_k}\\ \varphi_{u_k}\\ \varphi_{\rho_k}\end{bmatrix} \qquad \varphi_{0_k} \triangleq \begin{bmatrix}\varphi_{y_{0,k}}\\ \varphi_{u_{0,k}}\\ \varphi_{\rho_{0,k}}\end{bmatrix} \qquad \tilde{\varphi}_k \triangleq \begin{bmatrix}\tilde{\varphi}_{y_k}\\ \tilde{\varphi}_{u_k}\\ \tilde{\varphi}_{\rho_k}\end{bmatrix} \qquad (7)$$

where

$$\varphi_{y_k} \triangleq \begin{bmatrix}-y_{k-1}\\ \vdots\\ -y_{k-n_a}\end{bmatrix} \quad \varphi_{u_k} \triangleq \begin{bmatrix}u_{k-1}\\ \vdots\\ u_{k-n_b}\end{bmatrix} \quad \varphi_{\rho_k} \triangleq \begin{bmatrix}y_{k-1}u_{k-1}\\ \vdots\\ y_{k-p}u_{k-p}\end{bmatrix}$$

$$\varphi_{y_{0,k}} \triangleq \begin{bmatrix} -y_{0_{k-1}} \\ \vdots \\ -y_{0_{k-n_a}} \end{bmatrix} \quad \varphi_{u_{0,k}} \triangleq \begin{bmatrix} u_{0_{k-1}} \\ \vdots \\ u_{0_{k-n_b}} \end{bmatrix} \quad \varphi_{\rho_{0,k}} \triangleq \begin{bmatrix} y_{0_{k-1}} u_{0_{k-1}} \\ \vdots \\ y_{0_{k-p}} u_{0_{k-p}} \end{bmatrix}$$

$$\tilde{\varphi}_{y_k} \triangleq \begin{bmatrix} -\tilde{y}_{k-1} \\ \vdots \\ \tilde{y}_{k-n_a} \end{bmatrix} \quad \tilde{\varphi}_{u_k}^T \triangleq \begin{bmatrix} \tilde{u}_{k-1} \\ \vdots \\ \tilde{u}_{k-n_b} \end{bmatrix} \quad \tilde{\varphi}_{\rho_k}^T \triangleq \begin{bmatrix} \tilde{\rho}_{k-1,k-1} \\ \vdots \\ \tilde{\rho}_{k-p,k-p} \end{bmatrix}$$

with $\varphi_k$, $\varphi_{0_k}$, $\tilde{\varphi}_k \in \mathcal{R}^{n_\theta}$, $\varphi_{y_k}$, $\varphi_{y_{0,k}}$, $\tilde{\varphi}_{y_k} \in \mathcal{R}^{n_a}$, $\varphi_{u_k}$, $\varphi_{u_{0,k}}$, $\tilde{\varphi}_{u_k} \in \mathcal{R}^{n_b}$, $\varphi_{\rho_k}$, $\varphi_{\rho_{0,k}}$, $\tilde{\varphi}_{\rho_k} \in \mathcal{R}^{n_\eta}$ and $\tilde{\rho}_{k-i,k-j}$ denoting the noise contribution corresponding to the bilinear product terms of the regressor vector $\varphi_{\rho_k}$. In agreement with (7), the extended regressor vectors are given by

$$\bar{\varphi}_k \triangleq \begin{bmatrix} -y_k \\ \varphi_k \end{bmatrix} \qquad \bar{\varphi}_{0_k} \triangleq \begin{bmatrix} -y_{0_k} \\ \varphi_{0_k} \end{bmatrix} \qquad \tilde{\bar{\varphi}}_k \triangleq \begin{bmatrix} -\tilde{y}_k \\ \tilde{\varphi}_k \end{bmatrix} \quad (8)$$

where $\bar{\varphi}_k$, $\bar{\varphi}_{0_k}$, $\tilde{\bar{\varphi}}_k \in \mathcal{R}^{n_\theta+1}$.

# 3 A BRIEF REVIEW OF BBCLS AND BFS

## 3.1 BBCLS

The BBCLS algorithm for the class of DBS comprises of equations (9), (10) and (11), see (Larkowski et al., 2007). These correspond to the bilinear bias compensation rule, the noise covariance matrix and the noise 'variance' of the bilinear terms, respectively, i.e.

$$\hat{\theta}^{BBCLS} \triangleq \left(\Sigma_{\varphi\varphi} - \Sigma_{\tilde{\varphi}\tilde{\varphi}}\right)^{-1} \Sigma_{\varphi y} \quad (9)$$

$$\Sigma_{\tilde{\varphi}\tilde{\varphi}} \triangleq \begin{bmatrix} \sigma_{\tilde{y}} I_{n_a} & 0 & 0 \\ 0 & \sigma_{\tilde{u}} I_{n_b} & 0 \\ 0 & 0 & \sigma_{\tilde{\rho}} I_{n_\eta} \end{bmatrix} \quad (10)$$

$$\sigma_{\tilde{\rho}} \triangleq \sigma_u \sigma_{\tilde{y}} + \sigma_y \sigma_{\tilde{u}} - \sigma_{\tilde{u}} \sigma_{\tilde{y}} \quad (11)$$

where $\sigma_u \triangleq E[u_k]$ and $\sigma_y \triangleq E[y_k]$ are the expected values of the measured system input and output signals and $\hat{(\cdot)}$ denotes an estimate. It is noted, that in the remainder of the paper, the expression $\Sigma_{ab}$ will be used as general notation for the correlation matrix of vectors $a_k$ and $b_k$, i.e. $\Sigma_{ab} = E[a_k b_k^T]$. Equation (9) can be alternatively restated as

$$\hat{\theta}^{BBCLS} = \hat{\theta}^{LS} + \Sigma_{\varphi\varphi}^{-1} \Sigma_{\tilde{\varphi}\tilde{\varphi}} \hat{\theta}^{BBCLS} \quad (12)$$

where $\hat{\theta}^{LS}$ denotes the LS estimate. It is implied from the BBCLS algorithm that the knowledge regarding noise variances corrupting input/output of a system together with variances of measured input/output signals is sufficient to obtain unbiased estimates of the true system parameters.

## 3.2 BFS

The Frisch scheme is a technique that allows the direct estimation of the input/output noise variances together with the parameters of a system (Beghelli et al., 1990; Söderström, 2006). As consequence, the *a priori* knowledge regarding the values of $\sigma_{\tilde{u}}$ and $\sigma_{\tilde{y}}$ is not required leading to a wider practical applicability. The extension of the FS, in the framework of the BBCLS technique, for the class of DBS has been proposed in (Larkowski et al., 2008). Define the partitioned extended data covariance matrix

$$\Sigma_{\bar{\varphi}\bar{\varphi}} \triangleq \begin{bmatrix} \Sigma_{\bar{\varphi}_y\bar{\varphi}_y} & \Sigma_{\varphi_u\bar{\varphi}_y}^T & \Sigma_{\varphi_\rho\bar{\varphi}_y}^T \\ \Sigma_{\varphi_u\bar{\varphi}_y} & \Sigma_{\varphi_u\varphi_u} & \Sigma_{\varphi_\rho\varphi_u}^T \\ \Sigma_{\varphi_\rho\bar{\varphi}_y} & \Sigma_{\varphi_\rho\varphi_u} & \Sigma_{\varphi_\rho\varphi_\rho} \end{bmatrix} \quad (13)$$

where $\Sigma_{\bar{\varphi}_y\bar{\varphi}_y} \in \mathcal{R}^{(n_a+1)\times(n_a+1)}$, $\Sigma_{\varphi_u\bar{\varphi}_y} \in \mathcal{R}^{n_b\times(n_a+1)}$, $\Sigma_{\varphi_u\varphi_u} \in \mathcal{R}^{n_b\times n_b}$, $\Sigma_{\varphi_\rho\bar{\varphi}_y} \in \mathcal{R}^{n_\eta\times(n_a+1)}$, $\Sigma_{\varphi_\rho\varphi_u} \in \mathcal{R}^{n_\eta\times n_b}$ and $\Sigma_{\varphi_\rho\varphi_\rho} \in \mathcal{R}^{n_\eta\times n_\eta}$. The BFS consists of three main phases, i.e. calculation of the maximal admissible value for $\sigma_{\tilde{u}}$, denoted $\sigma_{\tilde{u}}^{max}$ (14), determination of a functional relationship between $\sigma_{\tilde{y}}$ and $\sigma_{\tilde{u}}$ (16a) and the specification of a cost function to find a unique value of $\sigma_{\tilde{u}}$ (18a). The quantity $\sigma_{\tilde{u}}^{max}$ is given by:

$$\sigma_{\tilde{u}}^{max} \triangleq \lambda_{\min}(A_1^*) \quad (14)$$

where $\lambda_{\min}(A_1^*)$ denotes the least eigenvalue of the matrix $A_1^*$, and $\max(\cdot)$ is the maximum operator. The matrix $A_1^*$ is defined as:

$$A_1^* \triangleq A_1 - B_1 \Sigma_{\bar{\varphi}_y\bar{\varphi}_y}^{-1} B_1^T \quad (15a)$$

with

$$A_1 \triangleq \begin{bmatrix} \Sigma_{\varphi_u\varphi_u} & \Sigma_{\varphi_\rho\varphi_u}^T \\ \Sigma_{\varphi_\rho\varphi_u} & \Sigma_{\varphi_\rho\varphi_\rho} \end{bmatrix} \qquad B_1 \triangleq \begin{bmatrix} \Sigma_{\varphi_u\bar{\varphi}_y} \\ \Sigma_{\varphi_\rho\bar{\varphi}_y} \end{bmatrix} \quad (15b)$$

The functional relationship between $\sigma_{\tilde{y}}$ and $\sigma_{\tilde{u}}$ is described by

$$\sigma_{\tilde{y}} \triangleq \lambda_{\min}(A_2^*) \quad (16a)$$

where

$$A_2^* \triangleq A_2 - B_2 (\Sigma_{\varphi_u\varphi_u} - \sigma_{\tilde{u}} I_{n_b})^{-1} B_2^T \quad (16b)$$

with

$$A_2 \triangleq \begin{bmatrix} \Sigma_{\bar{\varphi}_y\bar{\varphi}_y} & \Sigma_{\varphi_\rho\bar{\varphi}_y}^T \\ \Sigma_{\varphi_\rho\bar{\varphi}_y} & \Sigma_{\varphi_\rho\varphi_\rho} - \sigma_y \sigma_{\tilde{u}} I_{n_\eta} \end{bmatrix} \quad B_2 \triangleq \begin{bmatrix} \Sigma_{\varphi_u\bar{\varphi}_y}^T \\ \Sigma_{\varphi_\rho\varphi_u} \end{bmatrix} \quad (16c)$$

The cost function utilized is based on the Yule-Walker equations, see (Diversi et al., 2006) for details. The instrumental vector (Söderström and Stoica, 1994) for the measured data is defined as:

$$\bar{\varphi}_k^{IV} \triangleq \bar{\varphi}_{k-n_a-1} \in \mathcal{R}^{n_{\theta+1}\times 1} \quad (17)$$

Using (17) the corresponding cost function is formulated as:

$$J(\hat{\bar{\theta}}) \triangleq \|\Sigma_{\bar{\varphi}^{IV}\bar{\varphi}} \hat{\bar{\theta}}\|_2^2 \quad (18a)$$

such that

$$J(\hat{\bar{\theta}}) = 0 \Leftrightarrow \hat{\bar{\theta}} = \bar{\theta} \quad (18b)$$

Table 1: Summary of the ERBFS algorithm.

| Step | Description | Procedure |
|---|---|---|
| 1 | Choose $\lambda_k$ and $j$ | $0 < \lambda_k < 1,\ j = 2n_a + 1$ |
| 2 | RLS initialization: | $\hat{\theta}^{LS}_{n_\theta} = 0,\ P_{n_\theta} = 10^3 I_{n_\theta},\ \hat{\sigma}^k_u = 0,\ \hat{\sigma}^k_y = 0$ |
| 2.1 | RLS loop start | **for** $k = n_\theta + 1 \dots j$ |
| 2.2 | Data weighting | $\gamma_k = 1/k$ |
| | Computation of: | |
| 2.3 | $L_k$, $\hat{\theta}^{LS}_k$, $P_k$ | $L_k = \frac{P_{k-1}\varphi_k}{\varphi_k^T P_{k-1}\varphi_k + \frac{1-\gamma_k}{\gamma_k}},\ \hat{\theta}^{LS}_k = \hat{\theta}^{LS}_{k-1} + L_k\left(y_k - \varphi_k^T \hat{\theta}^{LS}_{k-1}\right)$ |
| | | $P_k = \frac{1}{1-\gamma_k}\left(P_{k-1} - L_k \varphi_k^T P_{k-1}\right)$ |
| 2.4 | $\hat{\sigma}^k_u$ , $\hat{\sigma}^k_y$ | $\hat{\sigma}^k_u = \frac{k-1}{k}\hat{\sigma}^{k-1}_u + \frac{1}{k-1}u_k^2,\ \hat{\sigma}^k_y = \frac{k-1}{k}\hat{\sigma}^{k-1}_y + \frac{1}{k-1}y_k^2$ |
| 2.5 | $\Sigma^k_{\bar{\varphi}_\rho \bar{\varphi}_y}$ | $\Sigma^k_{\bar{\varphi}\bar{\varphi}} = \Sigma^{k-1}_{\bar{\varphi}\bar{\varphi}} + \gamma_k\left(\bar{\varphi}_k \bar{\varphi}_k^T - \Sigma^{k-1}_{\bar{\varphi}\bar{\varphi}}\right)$ |
| 2.6 | RLS loop end | **end** |
| 3 | BBCLS and BFS initialization | $\Sigma^k_{\bar{\varphi}^{IV}\bar{\varphi}} = 0,\ \hat{\sigma}^{max}_{\tilde{u}} = \lambda_{\min}(A^*_{1,j})$ |
| 3.1 | Recursive BBCLS start | **for** $k = j+1 \dots N$ |
| 3.2 | Data weighting | $\gamma_k = 1/k$ |
| 3.3 | Iterative BFS start | |
| | Computation of: | |
| 3.3.1 | $\Sigma^k_{\bar{\varphi}^{IV}\bar{\varphi}}$ | $\Sigma^k_{\bar{\varphi}^{IV}\bar{\varphi}} = \Sigma^{k-1}_{\bar{\varphi}^{IV}\bar{\varphi}} + \gamma_k\left(\bar{\varphi}_k^{IV}\bar{\varphi}_k^T - \Sigma^{k-1}_{\bar{\varphi}^{IV}\bar{\varphi}}\right)$ |
| 3.3.2 | $\hat{\sigma}^k_{\tilde{u}}$ | $\hat{\sigma}^k_{\tilde{u}} = \arg\min_{\hat{\sigma}^k_{\tilde{u}}} J(\hat{\theta}_k)$ |
| 3.3.3 | $\hat{\sigma}^k_u$ , $\hat{\sigma}^k_y$ | $\hat{\sigma}^k_u = \frac{k-1}{k}\hat{\sigma}^{k-1}_u + \frac{1}{k-1}u_k^2,\ \hat{\sigma}^k_y = \frac{k-1}{k}\hat{\sigma}^{k-1}_y + \frac{1}{k-1}y_k^2$ |
| 3.3.4 | $\Sigma^k_{\bar{\varphi}\bar{\varphi}}$ | $\Sigma^k_{\bar{\varphi}\bar{\varphi}} = \Sigma^{k-1}_{\bar{\varphi}\bar{\varphi}} + \gamma_k\left(\bar{\varphi}_k \bar{\varphi}_k^T - \Sigma^{k-1}_{\bar{\varphi}\bar{\varphi}}\right)$ |
| 3.3.5 | $A^k_2$, $B^k_2$ | $A^k_2 = \begin{bmatrix} \Sigma^k_{\bar{\varphi}_y\bar{\varphi}_y} & (\Sigma^k_{\bar{\varphi}_\rho\bar{\varphi}_y})^T \\ \Sigma^k_{\bar{\varphi}_\rho\bar{\varphi}_y} & \Sigma^T_{\varphi_\rho\varphi_\rho} - \hat{\sigma}^k_y\hat{\sigma}^k_{\tilde{u}} I_{n_\eta} \end{bmatrix},\ B^k_2 = \begin{bmatrix} (\Sigma^k_{\varphi_u\bar{\varphi}_y})^T \\ \Sigma^k_{\varphi_\rho\varphi_u} \end{bmatrix}$ |
| 3.3.6 | $A^*_{2,k}$ | $A^*_{2,k} = A^k_2 - B^k_2(\Sigma^k_{\varphi_u\varphi_u} - \hat{\sigma}^k_{\tilde{u}} I_{n_b})^{-1}(B^k_2)^T$ |
| 3.3.7 | $\hat{\sigma}^k_{\tilde{y}}$ | $\hat{\sigma}^k_{\tilde{y}} = \lambda_{\min}(A^*_{2,k})$ |
| 3.4 | Iterative BFS end | |
| 3.5 | $L_k$, $\hat{\theta}^{LS}_k$, $P_k$ | $L_k = \frac{P_{k-1}\varphi_k}{\varphi_k^T P_{k-1}\varphi_k + \frac{1-\gamma_k}{\gamma_k}},\ \hat{\theta}^{LS}_k = \hat{\theta}^{LS}_{k-1} + L_k\left(y_k - \varphi_k^T \hat{\theta}^{LS}_{k-1}\right)$ |
| | | $P_k = \frac{1}{1-\gamma_k}\left(P_{k-1} - L_k \varphi_k^T P_{k-1}\right)$ |
| 3.6 | $\hat{\sigma}^k_{\tilde{\rho}}$ | $\hat{\sigma}^k_{\tilde{\rho}} = \hat{\sigma}^k_u\hat{\sigma}^k_{\tilde{y}} + \hat{\sigma}^k_y\hat{\sigma}^k_{\tilde{u}} - \hat{\sigma}^k_{\tilde{u}}\hat{\sigma}^k_{\tilde{y}}$ |
| 3.7 | $\Sigma^k_{\tilde{\varphi}\tilde{\varphi}}$ | $\Sigma^k_{\tilde{\varphi}\tilde{\varphi}} = \begin{bmatrix} \hat{\sigma}^k_{\tilde{y}} I_{n_a} & 0 & 0 \\ 0 & \hat{\sigma}^k_{\tilde{u}} I_{n_b} & 0 \\ 0 & 0 & \hat{\sigma}^k_{\tilde{\rho}} I_{n_\eta} \end{bmatrix}$ |
| 3.8 | Bias compensation | $\hat{\theta}^{BBCLS}_k = \hat{\theta}^{LS}_k + P_k \Sigma^k_{\tilde{\varphi}\tilde{\varphi}} \hat{\theta}^{BBCLS}_{k-1}$ |
| 3.9 | Recursive BBCLS end | **end** |

# 4 RECURSIVE BBCLS WITH ITERATIVE BFS

In this section a recursive BBCLS (RBBCLS) algorithm is developed, which comprises the recursive update of the data and instrumental covariance matrices, whilst the bias compensation procedure is recursively applied to the data covariance matrix only. Furthermore, an application of the BFS at each iteration step is described with an additional extension incorporating the TR technique. It is to be noted that the RBBCLS method can be interpreted in the framework of the iterative bias compensating LS (BCLS) approaches, see e.g. (Zheng, 1998) and (Zheng, 2000) for more details.

### 4.1 Recursive BBCLS

The normalized recursive updates of the instrumental and data covariance matrices are given by the following equations, see (Ljung, 1999)

$$\Sigma^k_{\bar{\varphi}\varphi} = \Sigma^{k-1}_{\bar{\varphi}\varphi} + \gamma_k \left( \bar{\varphi}_k \varphi_k^T - \Sigma^{k-1}_{\bar{\varphi}\varphi} \right) \tag{19a}$$

$$\Sigma^k_{\bar{\varphi}^{IV}\bar{\varphi}} = \Sigma^{k-1}_{\bar{\varphi}^{IV}\bar{\varphi}} + \gamma_k \left( \bar{\varphi}_k^{IV} \varphi_k^T - \Sigma^{k-1}_{\bar{\varphi}^{IV}\bar{\varphi}} \right) \tag{19b}$$

with

$$\gamma_k = \left( \Sigma_{i=1}^k \beta_{k,i} \right)^{-1} = \frac{\gamma_{k-1}}{\lambda_k + \gamma_{k-1}} \tag{20}$$

where the *i*-th data is weighted at the discrete time *k* according to the following rule

$$\beta_{k,i} = \lambda_k \beta_{k-1,i} \qquad \text{for} \qquad 0 \le i \le k-1 \tag{21}$$

and $\beta_{k,k} = 1$. It is to be noted that (20) simplifies either to $1/k$ in the case of no adaptivity, i.e. when $\lambda_k = 1$ or to $1-\lambda$ when the exponential forgetting is used, i.e. $\lambda_k = \lambda$ with $0 < \lambda < 1$.

Assuming that the input/output noise variances, i.e. $\sigma_{\tilde{u}}$ and $\sigma_{\tilde{y}}$ are known *a priori* or can be estimated, allows application of the BBCLS algorithm to the recursively computed estimate of the parameter vector, i.e.

$$\hat{\theta}_k^{BBCLS} = \hat{\theta}_k^{LS} + P_k \Sigma_{\tilde{\varphi}\tilde{\varphi}} \hat{\theta}_{k-1}^{BBCLS} \tag{22a}$$

$$\Sigma_{\tilde{\varphi}\tilde{\varphi}} = \begin{bmatrix} \sigma_{\tilde{y}} I_{n_a} & 0 & 0 \\ 0 & \sigma_{\tilde{u}} I_{n_b} & 0 \\ 0 & 0 & \hat{\sigma}^k_{\tilde{\rho}} I_{n_\eta} \end{bmatrix} \tag{22b}$$

$$\hat{\sigma}^k_{\tilde{\rho}} = \hat{\sigma}^k_u \sigma_{\tilde{y}} + \hat{\sigma}^k_y \sigma_{\tilde{u}} - \sigma_{\tilde{u}} \sigma_{\tilde{y}} \tag{22c}$$

$$\hat{\sigma}^k_u = \frac{k-1}{k} \hat{\sigma}^{k-1}_u + \frac{1}{k-1} u_k^2 \tag{22d}$$

$$\hat{\sigma}^k_y = \frac{k-1}{k} \hat{\sigma}^{k-1}_y + \frac{1}{k-1} y_k^2 \tag{22e}$$

$$L_k = \frac{P_{k-1}\varphi_k}{\varphi_k^T P_{k-1} \varphi_k + \frac{1-\gamma_k}{\gamma_k}} \tag{22f}$$

$$\hat{\theta}_k^{LS} = \hat{\theta}_{k-1}^{LS} + L_k \left( y_k - \varphi_k^T \hat{\theta}_{k-1}^{LS} \right) \tag{22g}$$

$$P_k = \frac{1}{1-\gamma_k} \left( P_{k-1} - L_k \varphi_k^T P_{k-1} \right) \tag{22h}$$

It is remarked that whilst the input/output noise variances are postulated to be known, the noise 'variance' corresponding to the bilinear terms (22c) is required to be recursively approximated at each time step. This involves the recursive estimation of the variances of the input/output signals, i.e. (22d) and (22e) (Young, 1984). Note that due to assumptions A1 and A3, see (Pearson, 1999) for more details, the mean values of the input/output signals do not explicitly appear in expressions (22d) and (22e) since they are both null.

### 4.2 Iterative BFS

Utilization of the recursively evaluated data and instrumental covariance matrices from the RBBCLS algorithm allows the application of the BFS at each recursion. Thus it is possible not only to estimate the input/output noise variances but also to conduct the noise compensation procedure given by (22a), see (Linden et al., 2007) for more details regarding the linear case. This results in the ERBFS algorithm which is summarized in Table 1. It is to be noted that ERBFS is rather expensive from the computational point of view. However, with reference to Table 1, if the time allowed for the calculation of $\hat{\sigma}^k_{\tilde{u}}$ at step 3.3.2 is bounded, the algorithm at least satisfies the principles of a recursive estimation scheme (Ljung and Söderström, 1987; Ljung, 1999).

### 4.3 Regularized BFS

In the case when *a priori* knowledge regarding the value of the input noise variance is available, or can be approximately anticipated, a regularization technique may be utilized. The regularization method considered here is that of TR which forces the estimate towards a pre-specified value $\hat{\sigma}^*_{\tilde{u}}$ controlled by the parameter $\omega$ (Hansen, 2001). The incorporation of TR into the cost function given by (18a) results in the following regularized cost function

$$J(\hat{\bar{\theta}}, \omega, \hat{\sigma}^*_{\tilde{u}}) \triangleq \omega \| \Sigma_{\bar{\varphi}^{IV}\bar{\varphi}} \hat{\bar{\theta}} \|_2^2 + (1-\omega) \| \hat{\sigma}^*_{\tilde{u}} - \hat{\sigma}_{\tilde{u}} \|_2^2 \tag{23}$$

Note that (23) reduces to (18a) for $\omega = 1$. Furthermore, it may be beneficial to consider a variable controlling parameter, i.e. $\omega_k$, such that the impact of the regularization is significant at the beginning of the identification procedure but gradually diminishes as a function of the incoming data stream. It is proposed to realize the concept as follows

$$\omega_k = e^{-\frac{\varsigma}{k}} \tag{24}$$

where $\varsigma$ is a user defined parameter describing the rate at which the impact of the regularization diminishes. This choice allows the potential bias introduced by regularization (Hansen, 2001) to be alleviated as time progresses. With reference to Table 1, the introduction of regularization requires the additional setting of the parameter $\varsigma$ at step 3, the subsequent implementation of equation (24) between steps 3.3.1 and 3.3.2 and replacement of the cost function $J_k(\hat{\bar{\theta}}_k)$ from 3.3.2 by $J(\hat{\bar{\theta}}_k, \omega_k, \hat{\sigma}^*_{\tilde{u}})$.

# 5 SIMULATION STUDIES

This section provides a numerical evaluation and comparison of the proposed ERBFS algorithm with the RLS and the off-line BFS. The SISO DBS system used in the simulations with $n_a = 2$, $n_b = n_\eta = 1$ is simulated for $N = 5000$. It is described by the following difference equation

$$y_{0_k} = 1.2y_{0_{k-2}} - 0.9y_{0_{k-1}} + 0.6u_{0_{k-1}} + 0.1y_{0_{k-1}}u_{0_{k-1}} \quad (25)$$

The input is generated by

$$u_{0_k} \sim \mathcal{N}(0, 0.5) \quad (26)$$

The variances of the input and output noises are selected as $\sigma_{\tilde{u}} = 0.05$ and $\sigma_{\tilde{y}} = 0.16$, respectively, in order to yield an approximately equal signal-to-noise ratio (SNR) on both input and output, i.e. $\mathrm{SNR}_u \approx \mathrm{SNR}_y \approx 10$[dB]. In the case of the ERBFS and the regularized ERBFS (RERBFS) the minimization procedure from step 3.3.2 (see Table 1) is restricted to a maximum of 10 iterations. The parameter $\lambda_k$ is set to unity, i.e. no adaptivity, for all evaluated algorithms.

Considering the results presented in Figure 2, where the ERBFS is compared with its off-line counterpart and the RLS, the following observations are made:

a) RLS yields estimates that are asymptotically biased for the case of all system parameters.

b) The estimate of the vector $\vartheta$ obtained via the off-line BFS is quite close to its true value.

c) The EBFS achieved virtually identical estimates as the off-line BFS at the last recursion step, i.e. $k = N$.

d) Estimates given by EBFS converge to their off-line counterparts obtained by the BFS algorithm over the succesive recursions.

e) There is a clear correlation between the quality of the estimated variances of the input/output signals and the quality of the estimated input/output noise variances which, subsequently, has an impact on the estimates of the system parameters.

f) The ERBFS encountered some difficulties in the estimation of the input noise variance in the initial part of the identification procedure, i.e. up to about first 1000 samples which is indicated by the relatively highly scattered values of $\hat{\sigma}_{\tilde{u}}$.

In fact observation (f) can be regarded as a premise for considering regularization of the input noise variance such that the uncertainty in the initial part of the identification is alleviated, leading to improved accuracy of the estimates. In the second experiment the ERBFS is compared with the RERBFS, where the parameters are set as follows: $\varsigma = N/\omega_0$ where $\omega_0 = 100$ and $\sigma^*_{\tilde{u}} = 1.5\sigma_{\tilde{u}}$. For completeness, the results obtained by the BFS are also included. Consideration of the results in Figure 3 leads to the following observations:

g) Although the guess of the regularization parameter $\sigma^*_{\tilde{u}}$ was rather 'rough', a substantial improvement w.r.t. the input noise variance is observed in the initial period of the recursion.

h) The impact of applying TR is also evident in the case of the estimated output noise variance and the system parameters leading to the faster convergence.

i) Due to the utilization of the exponential controlling variable weighting $\omega_k$ the results obtained by the RERBFS and ERBFS at $k = N$ are practically indistinguishable. As a consequence, any potential induced bias due to the use of regularization is kept to a minimum, for the case considered.

# 6 CONCLUSIONS

A new recursive technique, i.e. the RBBCLS method, for identification of the class of SISO DBS has been developed and evaluated. Within the RBBCLS framework the ERBFS algorithm has been formulated in which the Frisch equations are evaluated at each recursion. The further extension incorporating the variable regularization via the TR method, giving rise to the RERBFS, was considered and shown to be beneficial in the initial period of the identification procedure. The methods proposed have been demonstrated when applied to a SISO DBS EIV identification problem. Comparisons made with the standard RLS technique illustrates the superiority and relatively high noise robustness of the proposed algorithms.

The further work will address two outstanding issues. Firstly, the extension and subsequent recursive implementation of the BCLS method to a wider class of the polynomial nonlinear EIV systems. Secondly, the alleviation of the computational burden via the application of a fully recursive BFS based on gradient approaches and/or on other fast methods via linearization of the Frisch equations.

# REFERENCES

Beghelli, S., Guidorzi, R. P., and Soverini, U. (1990). The Frisch scheme in dynamic system identification. *Automatica*, 26(1):171–176.

Burnham, K. J. (1991). *Self-tuning Control for Bilinear Systems*. PhD thesis, Coventry Polytechnic.

Diversi, R., Guidorzi, R., and Soverini, U. (2006). Yule-Walker equations in the Frisch scheme solution of errors-in-variables identification problems. In *Proc. of the 17th Int. Symposium on Mathematical Theory of Networks and Systems*, Kyoto, Japan.

Ekman, M. (2005). *Modeling and Control of Bilinear Systems: Applications to the Activated Sludge Process*. PhD thesis, Uppsala University.

Hansen, P. C. (2001). Regularization tools: A matlab package for analysis and solution of discrete ill-posed problems. Technical report, Department of Mathematical Modelling, Technical University of Denmark.

Ikonen, E. and Najim, K. (2002). *Advanced Process Identification and Control*. Marcel Dekker, Inc., USA.

Kotta, U. and Nomm, S.and Zinober, A. (2003). On state space realizability of bilinear systems described by higher order difference equations. In *Proc. of 42nd IEEE Conf. on Decision and Control*, volume 6, pages 5685–5690.

Larkowski, T., Linden, J. G., Vinsonneau, B., and Burnham, K. J. (2008). A novel errors-in-variables approach for bilinear models: the bilinear Frisch scheme. Internal report no. CTAC/TL-1/2008, Control Theory and Applications Centre, Coventry University, Coventry.

Larkowski, T., Vinsonneau, B., and Burnham, K. J. (2007). Bilinear model identification in the errors-in-variables framework via the bias-compensating least squares. In *IAR and ACD Int. Conf.*, Grenoble, France.

Linden, G. J., Vinsonneau, B., and Burnham, K. J. (2007). Fast algorithms for recursive Frisch scheme system identification. In *IAR and ACD Int. Conf.*, Grenoble, France.

Liu, J. (1992). On stationarity and asymptotic inference of bilinear time series models. *Statistica Sinica*, 2:479–494.

Ljung, L. (1999). *System Identification - Theory for the User*. Prentice Hall PTR, New Jersey, USA, 2nd edition edition.

Ljung, L. and Söderström, T. (1987). *Theory and practice of recursive identification*. MIT Press, Cambridge, UK.

Martineau, S., Burnham, K. J., Haas, O. C. L., Andrews, G., and Heeley, A. (2004). Four-term bilinear pid controller applied to an industrial furnace. *Control Engineering Practice*, 12(4):457–464.

Mohler, R. R. (1991). *Nonlinear Systems: Applications to Bilinear Control*, volume 2. Prentice Hall, Englewood Cliffs, NJ.

Mohler, R. R. and Khapalov, A. Y. (2000). Bilinear control and application to flexible a.c. transmission systems. *Journal of Optimization Theory and Applications*, 105(3):621–637.

Pearson, R. K. (1999). *Discrete-Time Dynamic Models*. Oxford University Press, New York, USA.

Rao, T. S. and Gabr, M. M. (1984). *An Introduction to Bispectral Analysis and Bilinear Time Series Models*. Springer-Verlag, Berlin, Germany.

Söderström, T. (2006). Statistical analysis of the Frisch scheme for identifying errors-in-variables systems. Technical report, Upsala Univercity, Department of Information Technology, Upsala, Sweden.

Söderström, T. (2007). Errors-in-variables methods in system identification. In *Automatica*, volume 43, pages 939–958.

Söderström, T., Soverini, U., and Mahata, K. (2002). Perspectives on errors-in-variables estimation for dynamic systems. In *Signal Processing*, volume 82(8), pages 1139–1154.

Söderström, T. and Stoica, P. (1994). *System Identification*. Prentice Hall Int., New Jersey, USA.

Young, P. (1984). *Recursive Estimation and Time-Series Analysis*. Springer-Verlag, Berlin, Germany.

Yu, D. (1996). *Fault diagnosis for industrial systems with emphasis on bilinear systems*. PhD thesis, Coventry University.

Zheng, W. X. (1998). Transfer function estimation from noisy input and output data. *Int. Journal of Adaptive Control and Signal Processing*, 12:365–380.

Zheng, W. X. (2000). Parametric identification of linear noisy input-output systems. *Cybernetics and Systems*, 31(7):803–816.

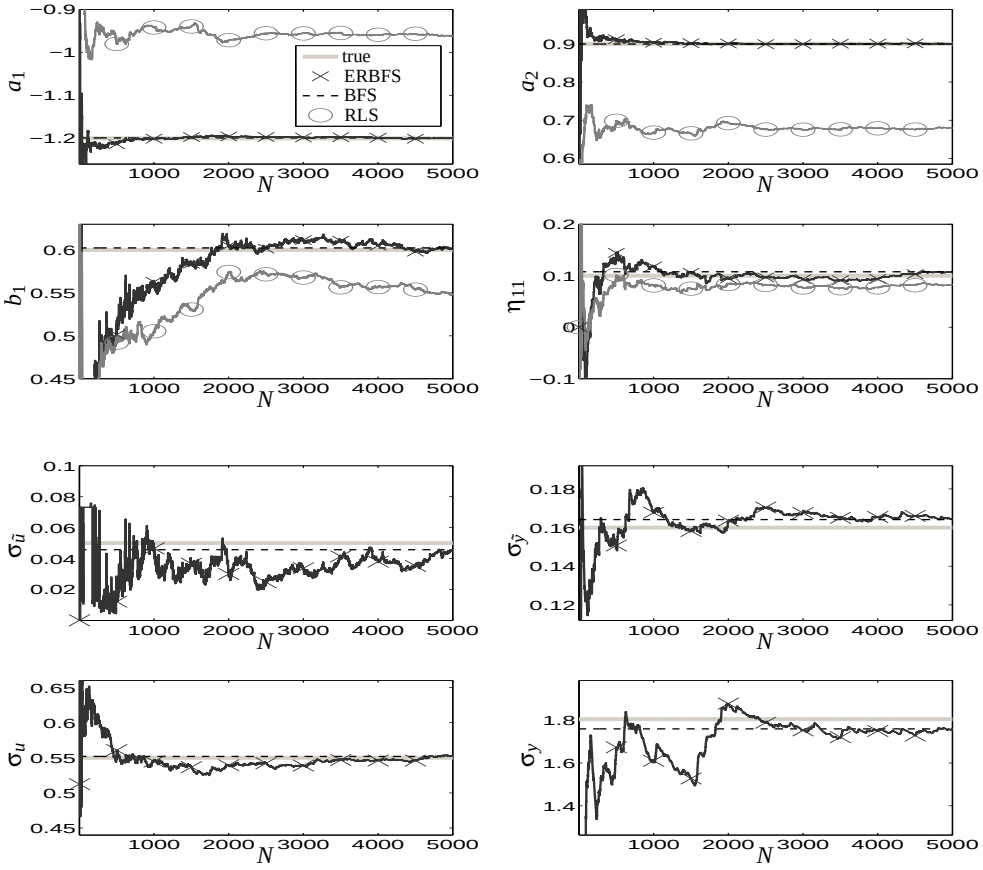



Figure 2: The results of the identification procedure using ERBFS, BFS and RLS algorithms.

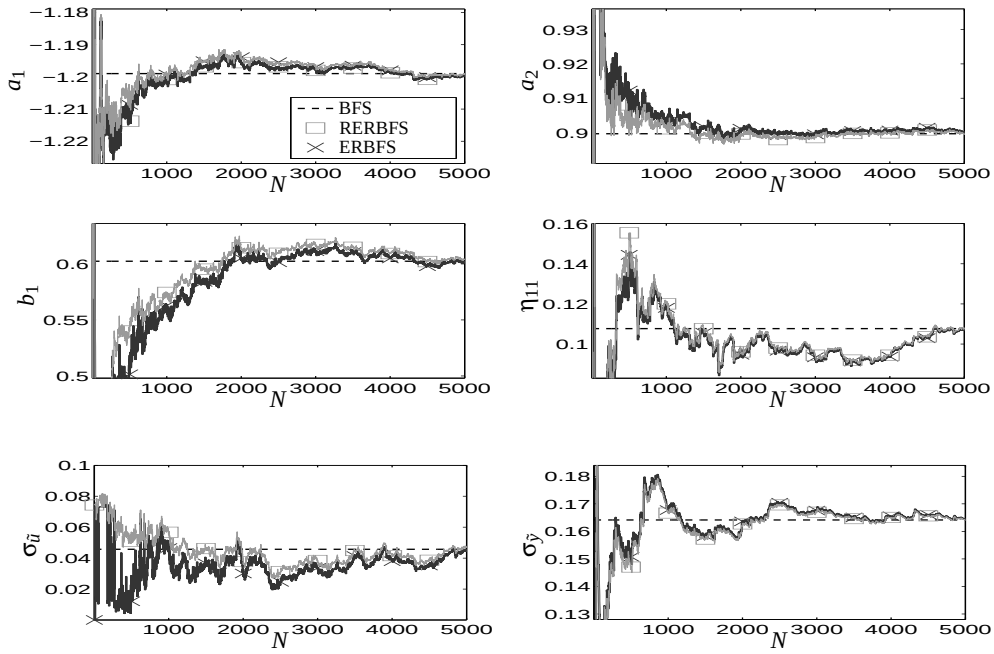



Figure 3: The results of the identification procedure using BFS, RERBFS and ERBFS algorithms.

# NEAR OPTIMUM CONTROL OF A FULL CAR ACTIVE SUSPENSION SYSTEM

Paolo Lino and Bruno Maione
*Dipartimento di Elettrotecnica ed Elettronica, Politecnico di Bari, via Re David 200, 70125, Bari, Italy*
*lino@deemail.poliba.it, maione@poliba.it*

Keywords: Active suspension, Suspension control, Virtual prototyping, Near-optimum control, AMESim©.

Abstract: In this paper, a near-optimum control strategy applied to a full car model equipped with an active suspension system is presented. The control law is based on a reduced order model obtained by means of a modal aggregation method, achieving a compromise between computational effort in deriving the control law and system performances. To assess the controller performances, a virtual prototype of the suspension system is developed by using AMESim, an advanced fluid-mechanic developing tool. The virtual prototype could be assumed as a reliable model of the real system enabling to perform safer and cheaper tests than using the real system. Simulation results show the effectiveness of the approach.

## 1 INTRODUCTION

A vehicle suspension system mainly aims to carry the car and its weight, control the vehicle direction of travel, keep the tires in contact with the road, and reduce the effect of shock forces due to road disturbances, braking and entries into curves. Handling and ride comfort can be significantly improved by using active suspension systems instead of passive or semi-active suspensions. More in details, passive suspension systems include spring and dampers characterized by static input-output relationships; semi-active suspensions use dampers with a variable damping coefficient; active suspensions apply a force on car body and wheels by means of an active actuator. The design process of control systems for active and semi-active suspensions is usually carried on by considering quarter car or half car models. The former only represents the vertical motion of the car body, the latter includes pitch or roll motions. Full car models give a more detailed representation of the car dynamics by including vertical displacement, pitch and roll dynamics at the same time. Different approaches to active suspensions control have been investigated by researchers, which are mainly based on fuzzy logic, adaptive control, LQR control and $H_\infty$ control, see (Yoshimura et al., 1997; Yoshimura et al., 1999; Huang and Lin, 2003; Al-Holou et al., 2002; Fialho and Balas, 2002; Alleyene and Hedrick, 1995; Hrovat, 1997) and the references therein.

The main drawback in using lower order models is that interactions between suspensions are neglected, so that the control action cannot compensate angular accelerations or sensibly improve stability. On the other hand, using simplified models makes the controller design easier. In this paper, a compromise between computational effort, detail in representing the system behaviour and controller performances is achieved by applying a near-optimum control strategy to a full car model equipped with an active suspension system. The controller performances are assessed by a virtual prototype of the suspension system. The virtual prototype could be assumed as a reliable model of the real system enabling to perform safer and cheaper tests than using the real system. Moreover, the integration of design and optimization processes of both mechanical and control subsystems are made easier by taking into account mutual interactions (Lino and Maione, 2007), thus reducing the whole design effort.

The proposed design process consists in few steps. Firstly, a $14^{th}$ order full car analytical model is developed, representing vertical car body and wheels motion, as well as pitch and roll angles dynamics. Then, a reduced order model is derived by applying a modal aggregation technique and used to develop a near-optimum control strategy. Finally, the virtual prototype of suspension system is built to validate the controller performances.

The paper is organized as follows. Sections 2 and 3 describe the full car analytical model and the virtual prototype of suspension system, respectively. The near-optimum control strategy is then introduced in Section 4. Some simulation results concerning the controlled system are shown in Section 5. Finally,

Section 6 gives some conclusions.

## 2 DYNAMICAL MODEL OF THE FULL CAR SUSPENSION SYSTEM

The full car model of a suspension system represents the vehicle as a rigid body with seven degrees of freedom, which originate from translation motion along axes, as well as rotational motions, i.e. pitch and roll motions around center of gravity (COG) (Ikenaga et al., 2000).

With reference to Fig. 1, by setting COG of the car body as origin of axes, the vehicle body dynamics can be described by the following equations:

$$\begin{cases} m_c\ddot{z} = f_{fl} + f_{fr} + f_{rl} + f_{rr} \\ I_{pc}\ddot{\theta} = -f_{fl}l_f + f_{fr}l_f + f_{rl}l_r + f_{rr}l_r \\ I_{rc}\ddot{\phi} = \frac{1}{2}f_{fl}Tr - \frac{1}{2}f_{fr}Tr + \frac{1}{2}f_{rl}Tr + \frac{1}{2}f_{rr}Tr \end{cases} \quad (1)$$

where $z$ is the vertical displacement of car body COG, $f_{fl}$, $f_{fr}$, $f_{rl}$, $f_{rr}$ are the forces applied by front-left, front-right, rear-left, rear-right suspensions on the car body, respectively, $\theta$ and $\phi$ are the pitch and roll angles, respectively, $l_f$ and $l_r$ are the distances from the front and rear axles to car body COG, $Tr$ is the wheels track, $m_c$ is the car body mass, $I_{pc}$ and $I_{rc}$ are the pitch and roll moments of inertia of the car body, respectively.

The force applied to the car body by each suspension can be computed as in the following:

$$\begin{cases} f_i = k_{s,i}(z_{w,i} - z_i) + c_{s,i}(\dot{z}_{w,i} - \dot{z}_i) + f_{A,i} \\ m_{w,i}\ddot{z}_{w,i} = -k_{s,i}(z_{w,i} - z_i) - c_{s,i}(\dot{z}_{w,i} - \dot{z}_i) + \\ \qquad - k_{w,i}(z_{w,i} - z_{r,i}) - f_{A,i} \end{cases} \quad (2)$$

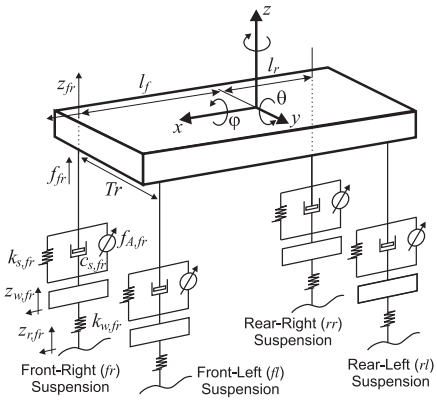



Figure 1: Vertical body model of the full car suspension system.

where subscript $i \in \{fl, fr, rl, rr\}$ characterizes the four suspensions, $z_{w,i}$ and $z_i$ are the vertical displacements of the wheel COG and car body corner, respectively, $m_{w,i}$ is the wheel mass, $k_{s,i}$ and $c_{s,i}$ are the stiffness and damping factor of the suspension, respectively, $k_{w,i}$ is the wheel stiffness, and $f_{A,i}$ is the active force applied by the controlled actuator.

The following equations express the vertical displacement of car body corners in terms of $z$, $\theta$ and $\phi$:

$$\begin{aligned} z_{fl} &= z + \frac{1}{2}Tr\cdot\sin\phi - l_f\sin\theta \approx z + \frac{1}{2}Tr\cdot\phi - l_f\theta \\ z_{fr} &= z - \frac{1}{2}Tr\cdot\sin\phi - l_f\sin\theta \approx z - \frac{1}{2}Tr\cdot\phi - l_f\theta \\ z_{rl} &= z + \frac{1}{2}Tr\cdot\sin\phi + l_r\sin\theta \approx z + \frac{1}{2}Tr\cdot\phi + l_r\theta \\ z_{rr} &= z - \frac{1}{2}Tr\cdot\sin\phi + l_r\sin\theta \approx z - \frac{1}{2}Tr\cdot\phi + l_r\theta \end{aligned} \quad (3)$$

where the approximations hold for small variations of pitch and roll angles. Combining equations (1), (2) and (3) results in a system of differential equations of the form:

$$\begin{cases} m_c\ddot{z}_c = f_1(z_c, \dot{z}_c, z_{w,i}, \dot{z}_{w,i}, \phi, \dot{\phi}, \theta, \dot{\theta}, f_{A,i}) \\ I_{pc}\ddot{\theta} = f_2(z_c, \dot{z}_c, z_{w,i}, \dot{z}_{w,i}, \phi, \dot{\phi}, \theta, \dot{\theta}, f_{A,i}) \\ I_{rc}\ddot{\phi} = f_3(z_c, \dot{z}_c, z_{w,i}, \dot{z}_{w,i}, \phi, \dot{\phi}, \theta, \dot{\theta}, f_{A,i}) \end{cases} \quad (4)$$

Equations (4), together with those of wheels vertical displacements:

$$m_r\ddot{z}_{w,i} = f_{4,i}(z_c, \dot{z_c}, z_{w,i}, \dot{z}_{w,i}, z_{r,i}, \theta, \dot{\theta}, \phi, \dot{\phi}, f_{A,i}) \quad (5)$$

represent the full car dynamics under the action of road disturbances and actuation forces. The design of a control law can be simplified by putting equations (4) and (5) in a state space form:

$$\dot{\mathbf{x}} = \mathbf{Ax} + \mathbf{Bf} + \mathbf{Hd}, \quad (6)$$

where $\mathbf{f} = [f_{A,fl}, f_{A,fr}, f_{A,rl}, f_{A,rr}]^T$ is the vector of actuation forces, $\mathbf{d} = [z_{r,fl}, z_{r,fr}, z_{r,rl}, z_{r,rr}]^T$ is the vector of road disturbances, and $\mathbf{x}$ is the state vector, composed of vertical displacements, pitch and roll angles and its derivatives.

A more accurate model of the suspension system includes the actuators dynamics. In this paper, the hydraulic actuator described in (Rajamani and Hedrick, 1995) is considered. It consists of a cylinder with a moving piston pushed by the pressure difference on its upper and lower surfaces (Fig. 2).

The pressure difference is regulated by an electronic valve driven by the control system. Under the assumption of a negligible piston inertia with respect to high hydraulic forces, the actuation force is given by:

$$f_{A,i} = -A_{y,i}(\dot{z}_{c,i} - \dot{z}_{w,i}) + u_i, \quad (7)$$

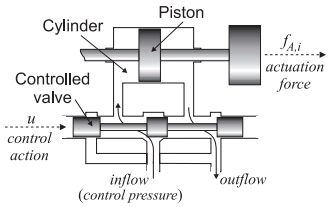



Figure 2: Hydraulic actuator for active suspensions.

where $u$ is a linear function of the control pressure set by acting on the electronic valve, and $A_{y,i}$ is a constant parameter depending on the system geometry and working fluid characteristics. By suitably introducing an $\mathbf{E}$ matrix, the system state space model becomes:

$$\dot{\mathbf{x}} = (\mathbf{A} + \mathbf{BE})\mathbf{x} + \mathbf{Bu} + \mathbf{Hd}, \tag{8}$$

where $\mathbf{u}$ is the vector of control inputs $u_i$.

# 3 VIRTUAL PROTOTYPE OF THE SUSPENSION SYSTEM

As virtual environment for design integration, we use AMESim© (Advanced Modelling Environment for Simulation): a simulation tool, which is oriented to lumped parameter modelling of components from different physical domains, interconnected by ports enlightening the energy exchanges between element and element and between an element and its environment. It also guarantees a flexible architecture, capable of including new components defined by the users (IMAGINE S.A., 2004).

The AMESim virtual prototype (Fig. 3) used to evaluate the controller performances has been developed by employing the AMESim-Simulink interface in Co-simulation mode: each suspension-wheel subsystem is modelled within the AMESim environment; the car body dynamical equations (1) are solved using MATLAB. AMESim and Simulink cooperate by integrating the relevant portions of models.

The main components of each suspension-wheel subsystem are the *Mass block with stiction and coulomb friction and end stops*, which computes the wheel dynamics through the Newtons second law of motion, the *Mechanical spring and dumper* computing the elastic and damping forces of suspensions and wheels depending on nonlinear stiffness and damping coefficients, the *Piston with moving body*, representing the actuator hydraulic circuit dynamics and computing the pressure forces acting upon the upper and lower piston surfaces, and the *3 positions hydraulic control valve* modelling the electro-hydraulic circuit driving the actuator.

The pressure dynamics inside cylinders are computed as a function of intake and outtake flows $Q_{in}$, $Q_{out}$, as well as of volume changes due to mechanical part motions, according to the following equation:

$$\frac{dP}{dt} = \frac{K_f}{v}\left(\rho\frac{dv}{dt} - Q_{in} + Q_{out}\right), \tag{9}$$

where $P$ and $\rho$ are the working fluid pressure and density, respectively, and $v$ is the taken up volume. $Q_{in}$ and $Q_{out}$ can be calculated by applying the energy conservation law, which gives, for a generic $Q$:

$$Q = c_D(\rho,\eta)A\rho\sqrt{\frac{2|\Delta P|}{\rho}}sgn(\Delta P), \tag{10}$$

where $\Delta P$ is the working fluid pressure difference across the flow section $A$; $sgn(\Delta P)$ is the sign function affecting the flow direction; the discharge coefficient $c_D$ accounts for nonuniform flow rates and flow process non-isentropicity, depending on fluid density $\rho$ and cinematic viscosity $\eta$. Finally, the *3 positions hydraulic control valve* block models the controlled valve as a second order spring-damp linear system.

To take into account the influence of inertia on pitch and roll dynamics during brakes and entries into a curve, the following equations are included in the model:

$$\begin{cases} I_{PC}\theta = m_c h_{cg}\ddot{x} \\ I_{PC}\phi = m_c h_{cg}\ddot{y} \end{cases}, \tag{11}$$

where $h_{cg}$ is the distance from the contact point between wheel and suspension to car body COG, and $\ddot{x}$ and $\ddot{y}$ are the COG accelerations along $x$ and $y$ axes, respectively.

# 4 THE NEAR-OPTIMUM CONTROL STRATEGY

Given a system described by the state space equations:

$$\begin{aligned} &\dot{\mathbf{x}}(t) = \mathbf{Ax}(t) + \mathbf{Bu}(t), \quad \mathbf{x}(0) = \mathbf{x}_0, \\ &\mathbf{y}(t) = \mathbf{Dx}(t) \end{aligned} \tag{12}$$

and a quadratic cost function:

$$J = \int_0^\infty \left(\mathbf{x}(t)^\mathbf{T}\mathbf{Q}\mathbf{x}(t) + \mathbf{u}(t)^\mathbf{T}\mathbf{R}\mathbf{u}(t)\right)dt, \tag{13}$$

being $\mathbf{Q}$ and $\mathbf{R}$ two positive semi-definite matrices, it is well known that the Linear-Quadratic-Regulator (LQR) problem consists in finding an input vector $\mathbf{u}^*(t) = -\mathbf{Kx}(t)$ minimizing the cost function $J$ via state-feedback (Dorato et al., 2000). The state feedback matrix $\mathbf{K}$ can be computed as:

$$\mathbf{K} = \mathbf{R}^{-1}\mathbf{B}^\mathbf{T}\mathbf{P}, \tag{14}$$

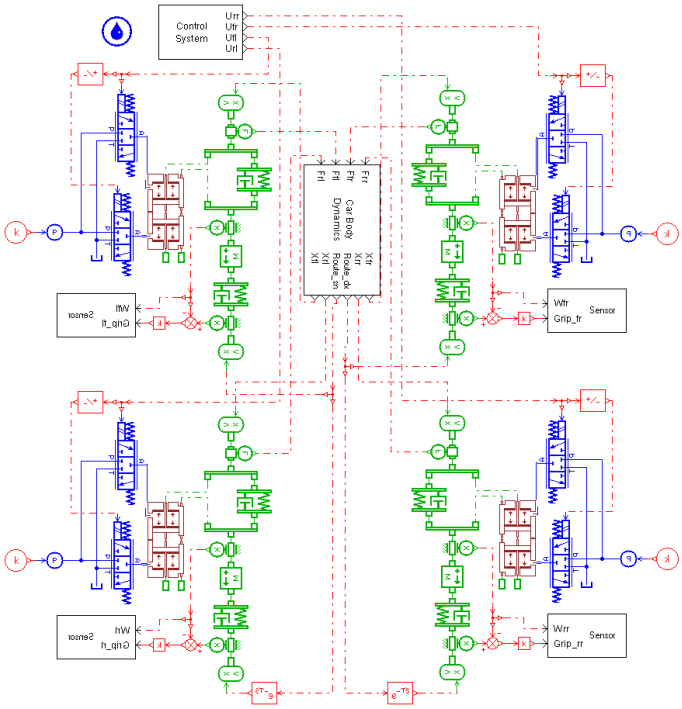



Figure 3: AMESim virtual prototype of the full car suspension system.

where $\mathbf{P}$ is the solution of the following matrix Riccati equation:

$$\mathbf{A}^{\mathbf{T}}\mathbf{P}+\mathbf{P}\mathbf{A}-\mathbf{P}\mathbf{B}\mathbf{R}^{-1}\mathbf{B}^{\mathbf{T}}\mathbf{P}+\mathbf{Q}=0. \qquad (15)$$

$\mathbf{Q}$ and $\mathbf{R}$ matrices can be considered as weights on state and control input, respectively, and affect the value assumed by elements of matrix $\mathbf{K}$. The Riccati equation complexity directly depends on the system order and its solution requests $n(n+1)/2$ operations. For a high order system it calls for a large computational effort, which could not be sustainable for on line calculations; a solution consists in adopting large scale system techniques to reduce problem complexity and make the application of optimal control theory easier (Jamshidi, 1983). These approaches, which are based on reduction of order, perturbation of parameters, decomposition of structure, hierarchical interaction or decentralization of control, lead to near optimality of system performance.

In this paper, the aggregation method based on the *modal approach* is applied to reduce the model order (Jamshidi, 1983). More in details, it neglects the effect of non dominant modes to obtain a system of *aggregated states* $\zeta$:

$$\begin{aligned}\dot{\zeta}(t)&=\mathbf{F}\zeta(t)+\mathbf{G}\mathbf{u}(t),\zeta(0)=\zeta_0\\ \dot{\mathbf{y}}(t)&=\mathbf{L}\zeta(t)\end{aligned} \qquad (16)$$

by using a transformation matrix $\mathbf{C}$:

$$\zeta(t)=\mathbf{C}\mathbf{x}(t),\quad \zeta(0)=\mathbf{C}\mathbf{x}(0). \qquad (17)$$

The reduced order model matches the full order model dynamics if the *dynamic exactness* condition holds:

$$\begin{aligned}\mathbf{F}\mathbf{C}&=\mathbf{C}\mathbf{A}\\ \mathbf{G}&=\mathbf{C}\mathbf{B}\\ \mathbf{L}\mathbf{C}&=\mathbf{D}.\end{aligned} \qquad (18)$$

By defining an error vector $\mathbf{e}(t)=\zeta(t)-\mathbf{C}\mathbf{x}(t)$, its dynamics is described by the equation $\dot{\mathbf{e}}=\mathbf{F}\mathbf{e}+(\mathbf{F}\mathbf{C}-\mathbf{C}\mathbf{A})\mathbf{x}+(\mathbf{G}-\mathbf{C}\mathbf{B})\mathbf{u}$, which reduces to $\dot{\mathbf{e}}=\mathbf{F}\mathbf{e}$ if conditions (18) hold. Provided that $\mathbf{F}$ is a positive definite matrix, error reduces to 0 even for $e(0)\neq 0$.

To derive the aggregation matrix $\mathbf{C}$, the modal approach exploits the system modal matrix $\mathbf{M}$, whose elements are the eigenvectors of state matrix $\mathbf{A}$. Hence, if arranging columns of matrix $\mathbf{M}$ by starting from eigenvectors related to slowest dynamics, the reduced order system matrices can be obtained using the following relationships:

$$\begin{aligned}\mathbf{F}&=\mathbf{M}_l\mathbf{S}\Lambda\mathbf{S}^{\mathbf{T}}\mathbf{M}_l^{-1}\\ \mathbf{C}&=\mathbf{M}_l\mathbf{S}\mathbf{M}^{-1}\\ \mathbf{G}&=\mathbf{C}\mathbf{B}\\ \mathbf{L}&\approx\mathbf{D}\mathbf{C}^{+}\end{aligned} \qquad (19)$$

where $\mathbf{M}_l$ is the nonsingular leading principal minor of matrix $\mathbf{M}$ of order $l$, $\mathbf{S}=[\mathbf{I}_l \quad 0]$, being $\mathbf{I}_l$ the Identity matrix, $\Lambda$ is the Jordan matrix of $\mathbf{A}$, and $\mathbf{C}^{+}$ is the pseudo-inverse of $\mathbf{C}$.

Considering the reduced order system, the Riccati equation becomes:

$$\mathbf{F}^{\mathbf{T}}\mathbf{P}_a + \mathbf{P}_{\mathbf{F}} - \mathbf{P}_a\mathbf{G}\mathbf{R}^{-1}\mathbf{G}^{\mathbf{T}}\mathbf{P}_a + \mathbf{Q}_a = 0, \tag{20}$$

so that the following control action is obtained:

$$\begin{aligned}\mathbf{u}_a(t) = -\mathbf{R}^{-1}\mathbf{G}^{\mathbf{T}}\mathbf{P}_a\zeta(t) = -\mathbf{R}^{-1}\mathbf{G}^{\mathbf{T}}\mathbf{P}_a\mathbf{C}\mathbf{x}(t) \\ = -\mathbf{K}_a\mathbf{x}(t)\end{aligned} \tag{21}$$

By pre- and post- multiplying eq. (20) by $\mathbf{C}^{\mathbf{T}}$ and $\mathbf{C}$, respectively, and considering the aggregation conditions, it is straightforward to obtain:

$$\begin{aligned}&\mathbf{A}^T(\mathbf{C}^{\mathbf{T}}\mathbf{P}_a\mathbf{C}) + (\mathbf{C}^{\mathbf{T}}\mathbf{P}_a\mathbf{C})\mathbf{A} + \\ &-(\mathbf{C}^{\mathbf{T}}\mathbf{P}_a\mathbf{C})\mathbf{B}\mathbf{R}^{-1}\mathbf{B}^{\mathbf{T}}(\mathbf{C}^{\mathbf{T}}\mathbf{P}_a\mathbf{C}) + \mathbf{C}^{\mathbf{T}}\mathbf{Q}_a\mathbf{C} = 0.\end{aligned} \tag{22}$$

Equations (15) and (22) coincide provided that the following positions hold:

$$\begin{aligned}\mathbf{P} &= \mathbf{C}^{\mathbf{T}}\mathbf{P}_a\mathbf{C}, \\ \mathbf{Q} &= \mathbf{C}^{\mathbf{T}}\mathbf{Q}_a\mathbf{C}.\end{aligned} \tag{23}$$

Hence, the $\mathbf{Q}_a$ matrix can be obtained as:

$$\mathbf{Q}_a = (\mathbf{C}\mathbf{C}^{\mathbf{T}})^{-1}\mathbf{C}\mathbf{Q}\mathbf{C}^{\mathbf{T}}(\mathbf{C}\mathbf{C}^{\mathbf{T}})^{-1}. \tag{24}$$

Finally, $\mathbf{K}_a$ matrix is obtained from $\mathbf{F}$, $\mathbf{G}$, $\mathbf{Q}_a$ and $\mathbf{R}$ matrices.

In this paper, a $6^{th}$ order aggregated model is derived from the full order suspension system and used to derive the control law.

## 5 SIMULATION RESULTS

To evaluate the controller performances, a set of tests is performed on the virtual prototype by applying different road profiles. In particular, the following benchmarks used in industrial practice are considered (Canale et al., 2006):

- *Sine wave hole* profile: a sine profile hole with 0.03 m of amplitude and 6 m of width; for the sake of brevity, only 30 and 90 Km/h car speeds are considered in this paper;
- *Short back* profile: a positive step of road profile with 0.02 m of amplitude and 0.5 m of width, with a travelling speed of 30 Km/h;
- *Drain well* profile: a negative step variation of road profile with 0.05 m of amplitude and 0.6 m of width, with a traveling speed of 30 km/h.

Moreover, the system response in case of entry in a curve or braking is analysed. The proposed controller performances have been compared with those obtained by using an active decoupled controller (Ikenaga et al., 2000) and an LQR controller based on the full order model. In particular, the control scheme proposed in (Ikenaga et al., 2000) includes a ride control loop, for road disturbances rejection, and an attitude control loop, for roll, pitch and vertical dynamics regulation. To former includes an active filtering feedback, the latter is based on a sky-hook control strategy. A decoupling is performed to deal with the under-actuation problem.

Figure 4 shows pitch and vertical displacement dynamics when applying the sine wave hole road profile. Since the road disturbance is symmetrically applied to left and right wheels, the roll dynamics is negligible and not displayed.

It is evident that the LQR controller guarantees better performances than the near-optimum and the decoupled controllers, in terms of overshoot and settling time. Nevertheless, the near optimum controller allows acceptable pitch angle and vertical displacement dynamics, improving the results obtained with the decoupled controller.

In Figure 5, the *drain well* and the *short back* disturbances are only applied to left wheels, so that the roll angle dynamics is excited. Simulation results show that the LQR controller still guarantees a better system behaviour for all conditions thanks to a prompt control action, while the near-optimum controller improves results obtained with the decoupled controller.

Finally, Figure 6 displays the effect of sudden longitudinal and lateral accelerations determined by braking (Fig. 6(a)), and entry into a curve (Fig. 6(b)), which independently affect the pitch dynamics and the roll dynamics, respectively.

In the former case, the near-optimum and the LQR controllers show similar performances; in the latter case, the near-optimum controller cannot reduce significantly the roll angle overshoot. In general, the decoupled controller cannot guarantee fast transients due to actuator saturation; the near optimum and LQR controller can suitably restrain the control action thanks to a suitable choice of the performance index weights. To sum up, the near optimum controller represents a compromise in terms of system performances and complexity, while the LQR controller always shows the best performances.

## 6 CONCLUSIONS

In this paper, a near-optimum control strategy applied to a full-car active suspension system has been proposed. The design process relies on the use of a high order analytical model, from which an aggregated low order model is derived, and of a virtual prototype developed by using the AMESim simulation package.

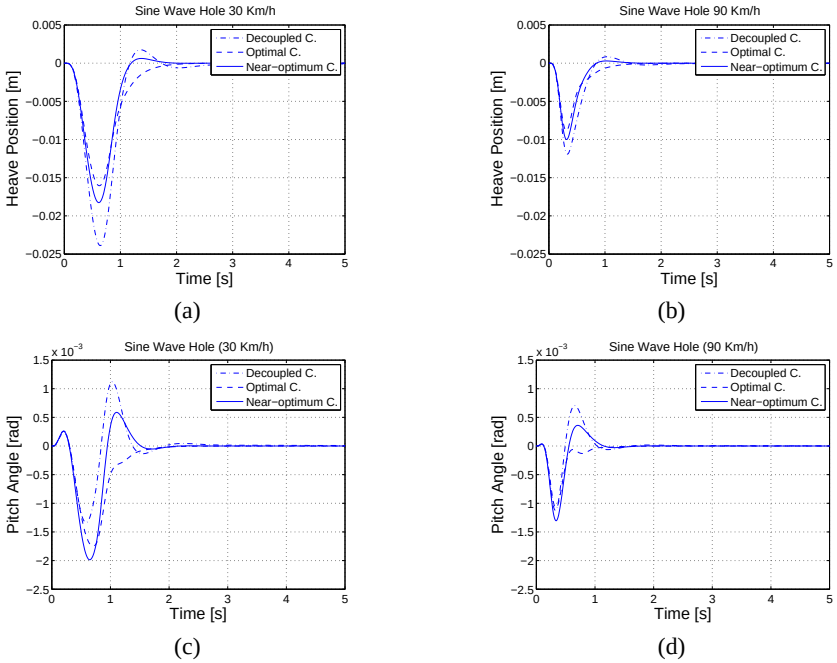



Figure 4: Vertical displacement (a)-(b) and Pitch angle (c)-(d) dynamics for a sine wave hole road profile ran at 30 km/h and 90 km/h, respectively.

The virtual prototype represents a reliable benchmark for evaluating the controller performances before the implementation on the real system. Moreover, it can be used for the integrated design of both mechanical and control subsystems at the same time. As simulation experiments have shown, the proposed controller provides good performances, despite the low computational effort required.

## REFERENCES


Al-Holou, N., Lahdhiri, T., Joo, D., Weaver, J., and Al-Abbas, F. (2002). Sliding mode neural network inference fuzzy logic control for active suspension systems. *IEEE Transactions on Fuzzy Systems*, 10(2):234–246.

Alleyene, A. and Hedrick, J. (1995). Nonlinear adaptive control of active suspensions. *IEEE Transactions on Control Systems Technology*, 3(1):94–101.

Canale, M., Milanese, M., and Novara, C. (2006). Semi-Active Suspension Control Using 'Fast' Model-Predictive Techniques. *IEEE Transactions on Control Systems Technology*, 14(6):1034–1046.

Dorato, P., Abdallah, C., and Cerone, V. (2000). *Linear Quadratic Control: An Introduction*. Krieger Publishing Company, Melbourne.

Fialho, I. and Balas, G. (2002). Road adaptive active suspension design using linear parameter-varying gain-scheduling. *IEEE Transactions on Control Systems Technology*, 10(1):43–54.

Hrovat, D. (1997). Survey of Advanced Suspension Developments and Related Optimal Control Applications. *Automatica*, 33(10):1781–1817.

Huang, S. and Lin, W. (2003). Adaptive fuzzy controller with sliding surface for vehicle suspension control. *IEEE Transactions on Fuzzy Systems*, 11(4):550–559.

Ikenaga, S., Lewis, F., Campos, J., and Davis, L. (2000). Active suspension control of ground vehicle based on a full-vehicle model. In *ACC 2000, Proceedings of the 2000 American Control Conference*.

Jamshidi, M. (1983). *Large-Scale Systems: Modeling and Control*. Elsevier Science Ltd, Amsterdam.

Lino, P. and Maione, B. (2007). Integrated design of a mechatronic system - the pressure control in common rails. In *ICINCO 2007, Proceedings of the Fourth International Conference on Informatics in Control, Automation and Robotics*, Angers, France.

Rajamani, R. and Hedrick, J. (1995). Adaptiver Observers for Active Automotive Suspensions: Theory and Experiment. *IEEE Transactions on Control Systems Technology*, 3(1):86–93.

IMAGINE S.A. (2004). *AMESim User Manual v4.2*. Roanne, France.

Yoshimura, T., Isari, Y., Li, Q., and Hino, J. (1997). Active suspension of motor coaches using skyhook damper and fuzzy logic control. *Control Engineering Practice*, 5(2):175–184.

Yoshimura, T., Nakaminami, K., Kurimoto, M., and Hino, J. (1999). Active suspension of passenger cars using linear and fuzzy-logic controls. *Control Engineering Practice*, (7):41–47.

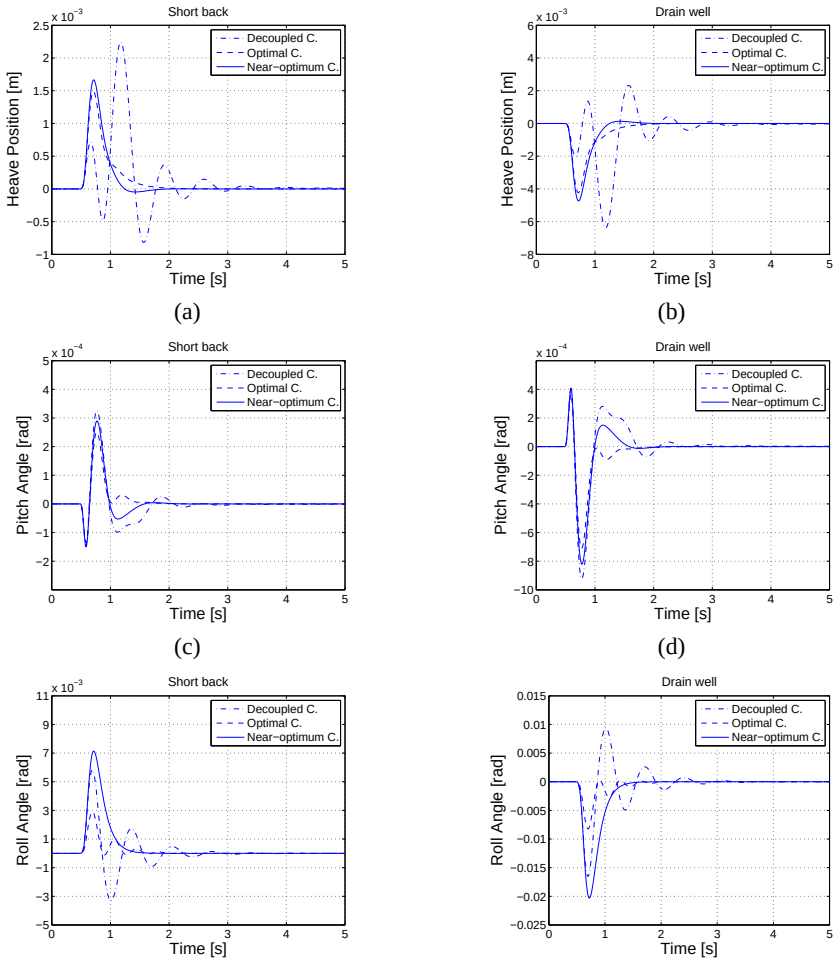



Figure 5: Vertical displacement, pitch angle and roll angle dynamics when applying *short back* and *drain well* road disturbances; (a)-(c)-(e) *short back* disturbance applied ; (b)-(d)-(f) drain well disturbance applied.

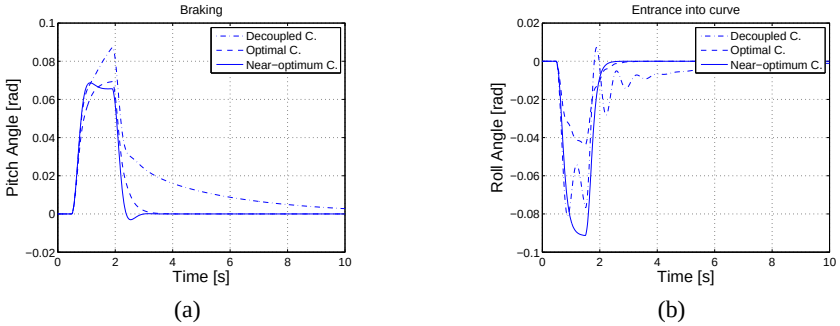



Figure 6: Pitch angle (a) and roll angle (b) dynamics in case of braking and entry into a curve, respectively.

# DESIGN AND IMPLEMENTATION OF A LOW-COST ATTITUDE AND HEADING NONLINEAR ESTIMATOR

Philippe Martin and Erwan Salaün
*Centre Automatique et Systèmes, École des Mines de Paris, 60 boulevard Saint-Michel, 75272 Paris Cedex 06, France*
*philippe.martin@ensmp.fr, erwan.salaun@ensmp.fr*

Keywords: Observers, sensor fusion, nonlinear filters, strapdown systems, invariance, inertial navigation, extended Kalman filters.

Abstract: In this paper we propose a nonlinear observer (i.e. a "filter") for estimating the orientation of a flying rigid body, using measurements from low-cost inertial and magnetic sensors. It has by design a nice geometrical structure appealing from an engineering viewpoint; it is easy to tune, computationally very economic, and with guaranteed (at least local) convergence around every trajectory. Moreover it behaves sensibly in the presence of acceleration and magnetic disturbances. We illustrate its good performance on experimental comparisons with a commercial system, and demonstrate its simplicity by implementing it on a 8-bit microcontroller.

## 1 INTRODUCTION

Aircraft, especially Unmanned Aerial Vehicles (UAV), commonly need to know their orientation to be operated, whether manually or with computer assistance. When cost or weight is an issue, using very accurate inertial sensors for "true" (i.e. based on the Schuler effect due to a non-flat rotating Earth) inertial navigation is excluded. Instead, low-cost systems –often called Attitude Heading Reference Systems (AHRS)– rely on light and cheap "strapdown" gyroscopes, accelerometers and magnetometers. The various measurements are "merged" according to the motion equations of the aircraft assuming a flat non-rotating Earth, usually with some kind of Extended Kalman Filter (EKF). For more details about avionics, various inertial navigation systems and sensor fusion, see for instance (Collinson, 2003; Kayton and Fried, 1997; Grewal et al., 2007) and the references therein.

While the EKF is a general method capable of good performance when properly tuned, it suffers several drawbacks: it is not easy to choose the numerous parameters; it is computationally expensive, which is a problem in low-cost embedded systems; it is usually difficult to prove the convergence, even at first-order, and the designer has to rely on extensive simulations.

An alternative route is to use an ad hoc nonlinear observer as proposed in (Thienel and Sanner, 2003; Mahony et al., 2005; Hamel and Mahony, 2006; Martin and Salaün, 2007; Mahony et al., 2008). In the absence of a general method the main difficulty is of course to find such an observer. In this paper we use the rich geometric structure of the attitude-heading problem to derive an observer by the method developed in (Bonnabel et al., 2007), building up on the preliminary work (Martin and Salaün, 2007). It has by design a nice geometrical structure appealing from an engineering viewpoint; it is easy to tune, computationally very economic, and with guaranteed (at least local) convergence around every trajectory. Moreover it behaves sensibly in the presence of acceleration and magnetic disturbances. We illustrate its good performance on experimental comparisons with the commercial Microbotics MIDG II system, and demonstrate its simplicity by implementing it on a 8-bit Atmel AVR microcontroller.

## 2 THE PHYSICAL SYSTEM

### 2.1 Motion Equations

The motion of a flying rigid body (assuming the Earth is flat and defines an inertial frame) is described by

$$\dot{q} = \frac{1}{2} q * \omega$$
$$\dot{V} = A + q * a * q^{-1},$$

where:

- $q$ is the quaternion representing the orientation of the body with respect to the Earth-fixed frame
- $\omega$ is the instantaneous angular velocity vector

- $V$ is the velocity vector of the center of mass with respect to the Earth-fixed frame
- $A = ge_3$ is the (constant) gravity vector in North-East-Down coordinates (the unit vectors $e_1, e_2, e_3$ point respectively North, East, Down)
- $a$ is the specific acceleration vector, in this case the aerodynamic forces divided by the body mass.

The first equation describes the kinematics of the body, the second is Newton's force law. It is customary to use quaternions instead of Euler angles since they provide a global parametrization of the body orientation, and are well-suited for calculations and computer simulations. For more details see (Stevens and Lewis, 2003) or any other good textbook on aircraft modeling, and section 7 for useful formulas used in this paper.

## 2.2 Measurements

We use three triaxial sensors providing nine scalar measurements: 3 gyros measure $\omega_m = \omega + \omega_b$, where $\omega_b$ is a constant vector bias; 3 magnetometers measure $y_B = b_s q^{-1} * B * q$, where $B = B_1 e_1 + B_3 e_3$ is the Earth magnetic field in NED coordinates and $b_s > 0$ is a constant scaling factor; 3 accelerometers measure $a_m = a_s a$, where $a_s > 0$ is a constant scaling factor. As customary in low-cost unaided attitude heading systems we assume the linear acceleration $\dot{V}$ small, hence approximate the accelerometers measurements by $y_A := a_s q^{-1} * A * q$ (the sign is reversed for convenience). All the nine measurements are of course also corrupted by noise.

There is some freedom in modeling the sensors imperfections. A simple first-order observability analysis reveals that up to six unknown constants can be estimated: besides the gyro bias $\omega_b$, it is possible to estimate two imperfections on $y_B$ and one on $a_m$, or one on $y_B$ and two on $a_m$. Nevertheless it is impossible to model three imperfections on $a_m$: in particular if we write $a_m = a + a_b$, with $a_b$ a constant vector bias, only two components of $a_b$ are observable; moreover, only one imperfection on $a_m$ can be estimated without relying on the possibly disturbed magnetic measurements. On the other hand it is also impossible to estimate the three components of the magnetic field $B$, but only the North and Down components.

In an AHRS it is usually desirable to use the magnetic measurements to estimate only the heading, so that a magnetic disturbance does not affect the estimated attitude. We will see this can be achieved by considering $y_C := y_A \times y_B = c_s q^{-1} * C * q$, where $c_s := a_s b_s > 0$ and $C := A \times B$, rather than the direct measurement $y_B$. Notice that $\langle y_A, y_C \rangle = \langle A, C \rangle = 0$, so that we are left with 8 independent measurements; as a consequence only five unknown constants can now be estimated. This is not a drawback and is even beneficial since the observer will then not depend on the latitude-varying $B_3$.

## 2.3 The Model

To design our observer we thus consider the system

$$\dot{q} = \frac{1}{2} q * (\omega_m - \omega_b) \tag{1}$$
$$\dot{\omega}_b = 0 \tag{2}$$
$$\dot{a}_s = 0 \tag{3}$$
$$\dot{c}_s = 0 \tag{4}$$

with the output

$$\begin{pmatrix} y_A \\ y_C \end{pmatrix} = \begin{pmatrix} a_s q^{-1} * A * q \\ c_s q^{-1} * C * q \end{pmatrix}. \tag{5}$$

This system is observable since all the state variables can be recovered from the known quantities $\omega_m, y_A, y_C$ and their derivatives: from (5), $a_s = \frac{1}{g} \|y_A\|$ and $c_s = \frac{1}{B_1 g} \|y_C\|$; hence we know the action of $q$ on the two independent vectors $A$ and $C$, which completely defines $q$ in function of $y_A, y_C, a_s, c_s$. Finally (1) yields $\omega_b = \omega_m - 2q^{-1}\dot{q}$.

# 3 THE NONLINEAR OBSERVER

## 3.1 Invariance of the System Equations

There is no general method for designing a nonlinear observer for a given system. When the system has a rich geometric structure the recent paper (Bonnabel et al., 2007) provides a constructive method to this problem. In short, when a system of state dimension $n$ is invariant by a transformation group of dimension $r \leq n$, the method yields the general invariant preobserver. More importantly the error between the estimated and actual states is described in suitable invariant coordinates by a system of dimension $2n - r$; in particular the error system has dimension $n$ when the group has dimension $n$. This result, which is reminiscent of the linear case, greatly simplifies the convergence analysis.

In our case the transformation generated by constant rotations and translations in the body-fixed frame and output scaling

$$\varphi_{(q_0,\omega_0,a_0,c_0)}\begin{pmatrix} q \\ \omega_b \\ a_s \\ c_s \end{pmatrix} = \begin{pmatrix} q * q_0 \\ q_0^{-1} * \omega_b * q_0 + \omega_0 \\ a_0 a_s \\ c_0 c_s \end{pmatrix}$$

$$\psi_{(q_0,\omega_0,a_0,c_0)}(\omega_m) = q_0^{-1} * \omega_m * q_0 + \omega_0$$

$$\rho_{(q_0,\omega_0,a_0,c_0)}\begin{pmatrix} y_A \\ y_C \end{pmatrix} = \begin{pmatrix} a_0 q_0^{-1} * y_A * q_0 \\ c_0 q_0^{-1} * y_C * q_0 \end{pmatrix},$$

where $q_0$ is a unit quaternion, $\omega_0$ a vector in $\mathbb{R}^3$ and $a_0, c_0 > 0$, is easily seen to be a transformation group with the same dimension as the system (1)–(4).

The system (1)–(4) is indeed invariant by the transformation group since

$$\dot{\overbrace{q * q_0}} = \dot{q} * q_0 = \frac{1}{2}(q * q_0) * \big((q_0^{-1} * \omega_m * q_0 + \omega_0) - (q_0^{-1} * \omega_b * q_0 + \omega_0)\big)$$

$$\dot{\overbrace{q_0^{-1} * \omega_b * q_0 + \omega_0}} = q_0^{-1} * \dot{\omega}_b * q_0 = 0$$

$$\dot{\overbrace{a_0 a_s}} = a_0 \dot{a}_s = 0$$

$$\dot{\overbrace{c_0 c_s}} = c_0 \dot{c}_s = 0.$$

Notice also that from a physical and engineering viewpoint, it is perfectly sensible for an observer using measurements expressed in the body-fixed frame not to be affected by the actual choice of coordinates, i.e. by a constant rotation in the body-fixed frame. Similarly, a translation of the gyro bias by a vector constant in the body-fixed frame and output scalings should not affect the observer. This is precisely what the method of (Bonnabel et al., 2007) achieves.

## 3.2 The General Invariant Observer

Following the theory developed in (Bonnabel et al., 2007), see also (Martin and Salaün, 2007) for details, the general invariant observer writes

$$\dot{\hat{q}} = \frac{1}{2}\hat{q} * (\omega_m - \hat{\omega}_b) + (LE) * \hat{q} \tag{6}$$

$$\dot{\hat{\omega}}_b = \hat{q}^{-1} * (ME) * \hat{q} \tag{7}$$

$$\dot{\hat{a}}_s = \hat{a}_s NE \tag{8}$$

$$\dot{\hat{c}}_s = \hat{c}_s OE. \tag{9}$$

Here the invariant output error $E$ is the $5 \times 1$ vector

$$E := \Big(\langle E_A, e_1\rangle, \langle E_A, e_2\rangle, \langle E_A, e_3\rangle, \langle E_C, e_1\rangle, \langle E_C, e_2\rangle\Big)^T$$

made up of the projections of the vectors

$$E_A := A - \frac{1}{\hat{a}_s}\hat{q} * y_A * \hat{q}^{-1}$$

$$E_C := C - \frac{1}{\hat{c}_s}\hat{q} * y_C * \hat{q}^{-1}.$$

Only 5 of the 6 possible projections are independent since $\langle A, C\rangle = 0$ and $\langle y_A, y_C\rangle = 0$ imply

$$\langle E_A, E_C\rangle = \langle A, E_C\rangle + \langle E_A, C\rangle;$$

$L, M$ are $3 \times 5$ matrices and $N, O$ are $1 \times 5$ matrices with entries possibly depending on the components of $E$ and of the complete invariant $I$ defined by

$$I := \hat{q} * (\omega_m - \hat{\omega}_b) * \hat{q}^{-1}.$$

It is easy to check this observer is invariant. Notice also the two built-in desirable geometric features: $\|\hat{q}(t)\| = \|\hat{q}(0)\| = 1$ since $LE$ is a vector of $\mathbb{R}^3$ (see section 7); $\hat{a}_s(t), \hat{c}_s(t) > 0$ provided $\hat{a}_s(0), \hat{c}_s(0) > 0$.

## 3.3 The Invariant Error System

Following (Bonnabel et al., 2007), an invariant state error is given by

$$\begin{pmatrix} \eta \\ \beta \\ \alpha \\ \gamma \end{pmatrix} = \begin{pmatrix} \hat{q} * q^{-1} \\ q * (\hat{\omega}_b - \omega_b) * q^{-1} \\ \frac{a_s}{\hat{a}_s} \\ \frac{c_s}{\hat{c}_s} \end{pmatrix}.$$

Therefore,

$$\dot{\eta} = \dot{\hat{q}} * q^{-1} - \hat{q} * (q^{-1} * \dot{q} * q^{-1}) = (LE) * \eta - \frac{1}{2}\eta * \beta$$

$$\begin{aligned} \dot{\beta} &= q * (\dot{\hat{\omega}}_b - \dot{\omega}_b) * q^{-1} + \dot{q} * (\hat{\omega}_b - \omega_b) * q^{-1} \\ &\quad - q * (\hat{\omega}_b - \omega_b) * q^{-1} * \dot{q} * q^{-1} \\ &= (\eta^{-1} * I * \eta) \times \beta + \eta^{-1} * (ME) * \eta \end{aligned}$$

$$\dot{\alpha} = -\frac{a_s \dot{\hat{a}}_s}{\hat{a}_s^2} = -\alpha NE$$

$$\dot{\gamma} = -\frac{c_s \dot{\hat{c}}_s}{\hat{c}_s^2} = -\gamma OE.$$

Since $E$ is obtained from

$$E_A = A - \frac{a_s}{\hat{a}_s}\hat{q} * (q^{-1} * A * q) * \hat{q}^{-1} = A - \alpha\eta * A * \eta^{-1}$$

$$E_C = C - \gamma\eta * C * \eta^{-1}$$

we find as expected that the error system

$$\dot{\eta} = (LE) * \eta - \frac{1}{2}\eta * \beta \tag{10}$$

$$\dot{\beta} = (\eta^{-1} * I * \eta) \times \beta + \eta^{-1} * (ME) * \eta \tag{11}$$

$$\dot{\alpha} = -\alpha NE \tag{12}$$

$$\dot{\gamma} = -\gamma OE \tag{13}$$

depends only on the invariant state error $(\eta, \beta, \alpha, \gamma)$ and the "free" known invariant $I$, but not on the trajectory of the observed system (1)–(4). This property greatly simplifies the convergence analysis of the observer.

The linearized error system around the no-error equilibrium point $(\overline{\eta},\overline{\beta},\overline{\alpha},\overline{\gamma}) = (1,0,1,1)$ then reads

$$\delta\dot{\eta} = L\delta E - \frac{1}{2}\delta\beta \tag{14}$$
$$\delta\dot{\beta} = I \times \delta\beta + M\delta E \tag{15}$$
$$\delta\dot{\alpha} = -N\delta E \tag{16}$$
$$\delta\dot{\gamma} = -O\delta E, \tag{17}$$

where $\delta E$ is the $5\times 1$ vector

$$\Big(\langle\delta E_A,e_1\rangle,\langle\delta E_A,e_2\rangle,\langle\delta E_A,e_3\rangle,\langle\delta E_C,e_1\rangle,\langle\delta E_C,e_2\rangle\Big)^T = g\big(-2\delta\eta_2,2\delta\eta_1,-\delta\alpha,2B_1\delta\eta_3,-B_1\delta\gamma\big)^T$$

made up from the projections of the vectors

$$\delta E_A = A*\delta\eta - \delta\eta*A - \delta\alpha A = 2A\times\delta\eta - \delta\alpha A$$
$$\delta E_C = 2C\times\delta\eta - \delta\gamma C.$$

# 4 DESIGN OF $L,M,N,O$

Up to now, we have only investigated the structure of the observer. We now must choose the gain matrices $L,M,N,O$ to meet the following requirements:

- the error must converge to zero, at least locally
- the local error behavior should be easily tunable, if possible with a clear physical interpretation
- the magnetic measurements should not affect the attitude estimate, but only the heading
- the behavior in the face of acceleration and/or magnetic disturbances should be sensible and understandable.

## 4.1 Local Design

It turns out that the previous requirements can easily be met *locally around every trajectory* by taking

$$L := \frac{1}{2g}\begin{pmatrix} 0 & -l_1 & 0 & 0 & 0\\ l_2 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & -\frac{1}{B_1}l_3 & 0\end{pmatrix}$$
$$M := \frac{1}{2g}\begin{pmatrix} 0 & m_1 & 0 & 0 & 0\\ -m_2 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & \frac{1}{B_1}m_3 & 0\end{pmatrix}$$
$$N := \frac{1}{g}\begin{pmatrix}0 & 0 & -n & 0 & 0\end{pmatrix}$$
$$O := \frac{1}{B_1 g}\begin{pmatrix}0 & 0 & 0 & 0 & -o\end{pmatrix}$$

for all constant $l_1,l_2,l_3,m_1,m_2,m_3,n,o > 0$. This will follow from the very simple form of the linearized error system (14)–(17). We insist that it is not usually obvious to come up with a similar convergence result for an EKF.

Indeed, (14)–(17) now reads

$$\delta\dot{\eta} = \mathrm{D_l}\delta\eta - \frac{1}{2}\delta\beta \tag{18}$$
$$\delta\dot{\beta} = \mathrm{D_m}\delta\eta + I\times\delta\beta \tag{19}$$
$$\delta\dot{\alpha} = -n\delta\alpha \tag{20}$$
$$\delta\dot{\gamma} = -o\delta\gamma \tag{21}$$

where

$$\mathrm{D_l} = \begin{pmatrix} -l_1 & 0 & 0\\ 0 & -l_2 & 0\\ 0 & 0 & -l_3\end{pmatrix}$$
$$\mathrm{D_m} = \begin{pmatrix} m_1 & 0 & 0\\ 0 & m_2 & 0\\ 0 & 0 & m_3\end{pmatrix}.$$

When $I = 0$ (i.e. the system is at rest) the system completely decouples into:

- the longitudinal subsystem
$$\begin{pmatrix}\delta\dot{\eta}_2\\ \delta\dot{\beta}_1\end{pmatrix} = \begin{pmatrix}-l_1 & -\frac{1}{2}\\ 0 & m_1\end{pmatrix}\begin{pmatrix}\delta\eta_2\\ \delta\beta_1\end{pmatrix}$$
- the lateral subsystem
$$\begin{pmatrix}\delta\dot{\eta}_1\\ \delta\dot{\beta}_2\end{pmatrix} = \begin{pmatrix}-l_2 & -\frac{1}{2}\\ 0 & m_2\end{pmatrix}\begin{pmatrix}\delta\eta_1\\ \delta\beta_2\end{pmatrix}$$
- the heading subsystem
$$\begin{pmatrix}\delta\dot{\eta}_3\\ \delta\dot{\beta}_3\end{pmatrix} = \begin{pmatrix}-l_3 & -\frac{1}{2}\\ 0 & m_3\end{pmatrix}\begin{pmatrix}\delta\eta_3\\ \delta\beta_3\end{pmatrix}$$
- the scaling subsystem
$$\delta\dot{\alpha} = -n\delta\alpha$$
$$\delta\dot{\gamma} = -o\delta\gamma.$$

When $I \neq 0$ the longitudinal, lateral and heading subsystems are slightly coupled by the biases errors $\delta\beta$.

We now prove $(\delta\eta,\delta\beta,\delta\alpha,\delta\gamma)\to(0,0,0,0)$ whatever $(l_1,l_2,l_3,m_1,m_2,m_3,n,o) > 0$. The scaling subsystem obviously converges. For the other variables we consider the Lyapunov function

$$V = \frac{l_1}{2}\delta\eta_1^2 + \frac{l_2}{2}\delta\eta_2^2 + \frac{l_3}{2}\delta\eta_3^2 + \frac{1}{4}\|\delta\beta\|^2.$$

Differentiating $V$ and using $\langle\delta\beta, I\times\delta\beta\rangle = 0$, we get

$$\dot{V} = -\left(l_1m_1\delta\eta_1^2 + l_2m_2\delta\eta_2^2 + l_3m_3\delta\eta_3^2\right) \le 0.$$

Since $V$ is bounded from below, this implies that $V(\delta\eta(t),\delta\beta(t))$ converges as $t\to\infty$. Since

$$\lim_{t\to\infty}\int_0^t \dot{V}(\delta\eta(\tau),\delta\beta(\tau))d\tau = \lim_{t\to\infty} V(\delta\eta(t),\delta\beta(t)) - V(\delta\eta(0),\delta\beta(0)),$$

we conclude $\lim_{t\to\infty}\int_0^t \dot{V}(\delta\eta(\tau),\delta\beta(\tau))d\tau$ exists and is finite. On the other hand, $\dot{V}\leq 0$ also implies

$$0\leq V(\delta\eta(t),\delta\beta(t))\leq V(\delta\eta(0),\delta\beta(0)).$$

Therefore $\delta\eta(t)$ and $\delta\beta(t)$ are bounded. Equation (18) implies that $\delta\dot{\eta}(t)$ is bounded too, and finally that $\ddot{V}$ is bounded. Hence $\dot{V}$ is uniformly continuous and by Barbalat's lemma

$$\dot{V}\to 0 \;\;\Rightarrow\;\; \delta\eta\to 0.$$

Integrating (18), we get

$$\begin{aligned}\int_0^t \delta\dot{\eta}(\tau)d\tau &= \delta\eta(t)-\delta\eta(0)\\ &= \int_0^t (D_l\delta\eta(\tau)-\frac{1}{2}\delta\beta(\tau))d\tau.\end{aligned}$$

Since $\delta\eta(t)\to 0$, it follows

$$\lim_{t\to\infty}\int_0^t (D_l\delta\eta(\tau)-\frac{1}{2}\delta\beta(\tau))d\tau=-\delta\eta(0).$$

We assume $I$ is bounded, which is physically sensible. Since $\delta\eta(t)$ and $\delta\beta(t)$ are bounded, $D_l\delta\dot{\eta}(t)-\frac{1}{2}\delta\dot{\beta}(t)$ is bounded too. Hence $D_l\delta\eta(t)-\frac{1}{2}\delta\beta(t)$ is uniformly continuous. Applying Barbalat's lemma once again yelds

$$\lim_{t\to\infty}(D_l\delta\eta(t)-\frac{1}{2}\delta\beta(t))=0.$$

Since $\delta\eta\to 0$, we conclude $\delta\beta\to 0$, which ends the proof.

### 4.2 Global Design

We now look for correction terms ensuring global convergence whereas preserving the previous nice first-order properties. It is useful to define the following vectors and error output:

$$\begin{aligned}D &= C\times A\\ y_D &= y_C\times y_A\\ E_D &= D-\frac{1}{\hat{a}_s\hat{c}_s}\hat{q}*y_D*\hat{q}^{-1}.\end{aligned}$$

If we define the matrix $L$ by

$$LE:=\frac{l_a}{g^2}A\times E_A+\frac{l_c}{(B_1g)^2}C\times E_C+\frac{l_d}{(B_1g^2)^2}D\times E_D$$

it is easy to see that the zero-order part of $L$ is the same as the constant $L$ of section 4.1 with

$$\begin{aligned}l_1 &= l_a+l_c\\ l_2 &= l_a+l_d\\ l_3 &= l_c+l_d.\end{aligned}$$

To find a candidate Lyapunov function we also select the matrices $(M,N,O)$ by

$$ME:=\sigma LE$$

$$NE:=n\big(\frac{l_a}{g^2}E_A^T(E_A-A)+\frac{l_d}{(B_1g^2)^2}E_D^T(E_D-D)\big)$$

$$OE:=o\big(\frac{l_c}{(B_1g)^2}E_C^T(E_C-C)+\frac{l_d}{(B_1g^2)^2}E_D^T(E_D-D)\big)$$

with $(l_a,l_c,l_d,\sigma,n,o)>0$. The positive function $V:=\frac{1}{\sigma}\|\beta\|^2+\frac{l_a}{g^2}\|E_A\|^2+\frac{l_c}{(B_1g)^2}\|E_C\|^2+\frac{l_d}{(B_1g^2)^2}\|E_D\|^2$ satisfies $\dot{V}\leq 0$. The convergence proof follows the main lines of (Hamel and Mahony, 2006), with added technicalities.

With these choices the first-order behavior is the same as in section 4.1. The difference is the existence of the proportional factor $\sigma$ between the $l_i's$, $m_i's$ coefficients. We now have only 4 independent gains for the lateral, longitudinal and heading subsystems instead of 6. We choose to fix first the natural frequencies of the lateral $(\omega_l)$, longitudinal $(\omega_L)$ and heading $(\omega_h)$ subsystems. There is only 1 parameter left to fix the free damping ratios.

## 5 EFFECTS OF DISTURBANCES

Two main disturbances may affect the model. When $\dot{V}\neq 0$, the accelerometers measure in fact $a_sq^{-1}*A^**q$ where $A^*:=-\dot{V}+A$. Magnetic disturbances will also change $B$ into some $B^*$. For simplicity we consider that $A^*$, $B^*$ are constant. The measured outputs now become

$$\begin{pmatrix}y_{A^*}\\ y_{C^*}\end{pmatrix}=\begin{pmatrix}a_sq^{-1}*A^**q\\ c_sq^{-1}*C^**q\end{pmatrix}.$$

The error system is unchanged but E is now the $5\times 1$ vector

$$E:=\Big(\langle E_A,e_1\rangle,\langle E_A,e_2\rangle,\langle E_A,e_3\rangle,\langle E_C,e_1\rangle,\langle E_C,e_2\rangle\Big)^T$$

made up of the projections of the vectors

$$\begin{aligned}E_A &:= A-\frac{1}{\hat{a}_s}\hat{q}*y_{A^*}*\hat{q}^{-1}\\ E_C &:= C-\frac{1}{\hat{c}_s}\hat{q}*y_{C^*}*\hat{q}^{-1}.\end{aligned}$$

Let us define the points $(\overline{\eta},\overline{\beta},\overline{\alpha},\overline{\gamma})$ as following

$$\begin{aligned}\overline{\beta} &= 0\\ \overline{\eta}*A^**\overline{\eta}^{-1} &= (0\;0\;\|A^*\|)\\ \overline{\eta}*C^**\overline{\eta}^{-1} &= (0\;\|C^*\|\;0)\\ \overline{\alpha}=\frac{\|A^*\|}{\|A\|} &\quad\text{and}\quad \overline{\gamma}=\frac{\|C^*\|}{\|C\|}\end{aligned}$$

Doing the frame rotation defined by $\overline{\eta}$ we can define the new variables

$$\tilde{\eta} = \eta * \overline{\eta}^{-1} \qquad \tilde{\beta} = \overline{\eta} * \beta * \overline{\eta}^{-1}$$
$$\tilde{\alpha} = \alpha\overline{\alpha} \qquad \tilde{\gamma} = \gamma\overline{\gamma}.$$

The error system with these new variables writes

$$\dot{\tilde{\eta}} = -\frac{1}{2}\tilde{\eta} * \tilde{\beta} + (\tilde{L}\tilde{E}) * \tilde{\eta}$$
$$\dot{\tilde{\beta}} = (\tilde{\eta}^{-1} * \tilde{I} * \tilde{\eta}) \times \tilde{\beta} + \tilde{\eta}^{-1} * (\tilde{M}\tilde{E}) * \tilde{\eta}$$
$$\dot{\tilde{\alpha}} = -\tilde{\alpha}\tilde{N}\tilde{E}$$
$$\dot{\tilde{\gamma}} = -\tilde{\gamma}\tilde{O}\tilde{E}$$

where the new output error $\tilde{E}$ is made up of the projections of the vectors

$$\tilde{E}_A = A - \tilde{\alpha}\tilde{\eta} * A * \tilde{\eta}^{-1}$$
$$\tilde{E}_C = C - \tilde{\alpha}\tilde{\eta} * C * \tilde{\eta}^{-1}.$$

So $(\tilde{\eta}, \tilde{\beta}, \tilde{\alpha}, \tilde{\gamma})$ verify the same error system as $(\eta, \beta, \alpha, \gamma)$. In the new frame $(A^*, C^*)$ play the same role as $(A, C)$. All the properties of the observer are therefore preserved.

An important case is when only the magnetic field is perturbed, where we consider $A$ and $C^* = (C_1^* \; C_2^* \; 0)$ (instead of $C = (0 \; gB_1 \; 0)$). Expliciting the new equilibrium point $(\overline{\eta}, \overline{\beta}, \overline{\alpha}, \overline{\gamma})$ of the error system it can be seen that

$$\overline{\phi} = \overline{\theta} = 0 \text{ and } \overline{\psi} = \arctan\frac{C_1^*}{C_2^*}$$
$$\overline{\beta} = \overline{\alpha} = 0 \text{ and } \overline{\gamma} = \frac{\|C^*\|}{\|C\|},$$

where $(\overline{\phi}, \overline{\theta}, \overline{\psi})$ are the Euler angles corresponding to $\overline{\eta}$. In particular only the yaw angle $\overline{\psi}$ and $\overline{\gamma}$ are affected by the magnetic disturbance.

# 6 EXPERIMENTAL VALIDATION

We now compare the behavior of our observer with the commercial Microbotics MIDG II system used in Vertical Gyro mode. The following results have been obtained with the observer

$$\dot{\hat{q}} = \frac{1}{2}\hat{q} * (\omega_m - \hat{\omega}_b) + (LE) * \hat{q} + k(1 - \|\hat{q}\|^2)\hat{q}$$
$$\dot{\hat{\omega}}_b = \hat{q}^{-1} * (ME) * \hat{q}$$
$$\dot{\hat{a}}_s = \hat{a}_s NE$$
$$\dot{\hat{c}}_s = \hat{c}_s OE$$

and the choice of matrices defined by the parameters below. The added term $k(1 - \|\hat{q}\|^2)\hat{q}$ is a well-known numerical trick to keep $\|\hat{q}\| = 1$. Notice this term is also invariant.

We feed the observer with the raw measurements from the MIDG II gyros, acceleros and magnetic sensors. The observer is implemented in Matlab Simulink and its values are compared to the MIDG II results (computed according to the user manual by some Kalman filter). In order to have similar behaviors, we have chosen

$$l_a = 6e-2 \qquad l_c = 1e-1 \qquad l_d = 6e-2$$
$$m_a = 3.2e-3 \qquad m_c = 5.3e-3 \qquad m_d = 3.2e-3$$
$$n = 0.25 \qquad o = 0.5.$$

## 6.1 Comparison with a Commercial Device (Figure 1)

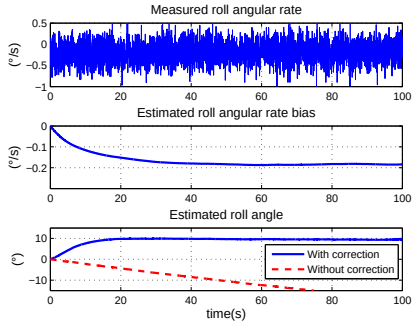
We first want to put in evidence the different properties of our observer we mentioned before. Therefore we did a long experience which can be divided into 3 parts:

- for $t < 240s$ the system is left at rest until the biases reach constant values. Fig.1(a) highlights the importance of the correction term in the angle estimation: without correction the estimated roll angle diverges with a slope of $-0.18^\circ/s$ (bottom plot), which is indeed the final value of the estimated bias (middle plot) (Fig.1(a) and 1(b)).
- for $240s < t < 293s$ we move the system in all directions. The observer and the MDG II give very similar results (Fig.1(c)).
- at $t = 385s$ the system is motionless and a magnet is put close to the sensors for 10s. As expected only the estimated yaw angle is affected by the magnetic disturbance (Fig.1(d)); the MIDG II exhibits a similar behavior.

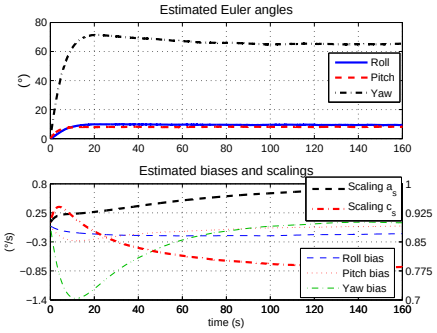
## 6.2 Influence of the Observer Correction Terms (Figure 2)

We have chosen the correction terms so as the magnetic measurements correct essentially the yaw angle and its corresponding bias, whereas the accelerometers measurements act on the other variables. We highlight this property on the following experiment (Fig.2). Once the biases have reached constant values, the system is left at rest during 35 minutes:

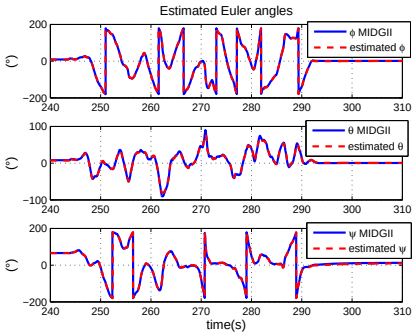
- for $t < 600s$ the results are very similar for the observer and the MIDG II.
- at $t = 600s$ the "magnetic correction terms" are switched off, i.e. the gains $l_c$,$l_d$,$m_c$,$m_d$ and $o$ are



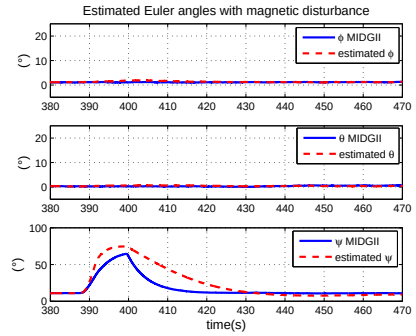

(a) Motionless estimated roll angle.




(b) Motionless estimated gyros biases and Euler angles.




(c) Dynamic estimated Euler angles.




(d) Estimated Euler angles with magnetic disturbances.

Figure 1: Experimental validation using Matlab.

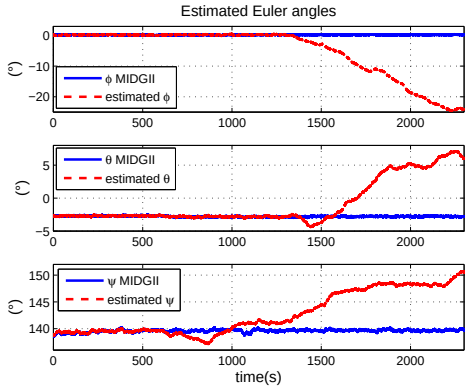



Figure 2: Influence of correction terms.

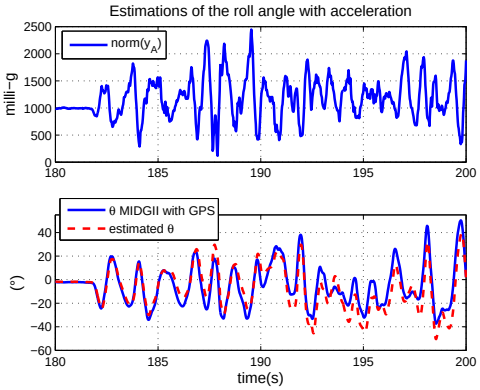



Figure 3: Estimated roll angle at $\dot{V} \neq 0$.

set to 0. The yaw angle estimated by the observer diverges because the corresponding bias is not perfectly estimated. Indeed, these variables are not observable without the magnetic measurements. The other variables are not affected.

- at $t = 1300s$ the "accelerometers correction terms" are also switched off, i.e. $l_a$,$m_a$ and $n$ are set to 0. All the estimated angles now diverge.

## 6.3 Acceleration Disturbance: $\dot{V} \neq 0$ (Figure 3)

The hypothesis $\dot{V} = 0$ may be wrong. In this case the observer does not converge any more to the right values. Indeed we illustrate this point on the figure 3 by comparing the roll angle estimated by our observer and the roll angle estimated by the MIDG II in INS mode (in this mode the attitude and heading estimations are aided by a GPS engine, hence are close to the "true" values).

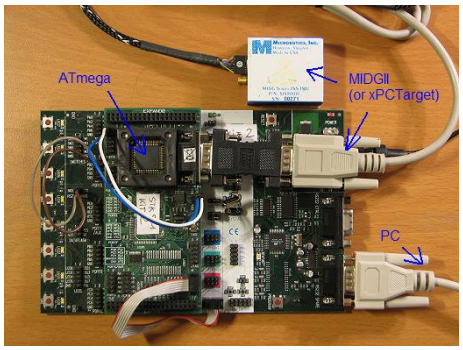



Figure 4: Experimental protocol.

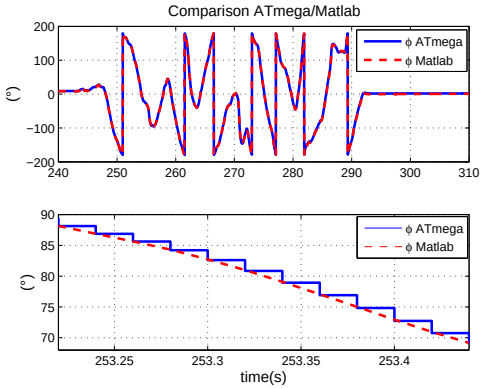



Figure 5: Comparison ATmega/Matlab.

## 7 IMPLEMENTATION ON A 8-BIT MICROCONTROLLER

In the preceding section we validated the Matlab Simulink implementation of our observer. To demonstrate its computational simplicity we have also implemented it on a 8-bit microcontroller (Atmel ATmega128 running at 11.0592MHz on development kit STK500/501). The computations are done in C with the standard floating point emulation. The microcontroller is fed with the MIDG II raw data at a 50Hz rate. We have used a simple Euler explicit approximation for the integration scheme.

The experimental protocol was the following:

1. move the sensors in all directions and save the MIDG II raw measurements and estimations at 50Hz.
2. use of xPCTarget to feed the ATmega with these data at 50Hz via a serial port and send at the same time the estimated variables. given by the microcontroller via a serial port to a computer
3. save these estimations.
4. compare offline the estimations given by the microcontroller and those given by the observer written in the Matlab embedded function.

This protocol can be illustrated by the figure 4 where xPCTarget has been replaced by the MIDG II.

We obtain the results of the figure 5. The two estimations are very similar. We see on the bottom plot the discretization at 50Hz due to the microcontroller.

## REFERENCES

Bonnabel, S., Martin, P., and Rouchon, P. (2007). Invariant observers. *arxiv.math.OC/0612193*. Accepted for publication in IEEE Trans. Automat. Control.

Collinson, R. (2003). *Introduction to avionics systems*. Kluwer Academic Publishers, second edition.

Grewal, M., Weill, L., and Andrews, A. (2007). *Global positioning systems, inertial navigation, and integration*. Wiley, second edition.

Hamel, T. and Mahony, R. (2006). Attitude estimation on $SO(3)$ based on direct inertial measurements. In *Proc. of the 2006 IEEE International Conference on Robotics and Automation*, pages 2170–2175.

Kayton, M. and Fried, W., editors (1997). *Avionics navigation systems*. Wiley, second edition.

Mahony, R., Hamel, T., and Pfimlin, J.-M. (2008). Nonlinear complementary filters on the special orthogonal group. *IEEE Trans. Automat. Control*. To appear.

Mahony, R., Hamel, T., and Pflimlin, J.-M. (2005). Complementary filter design on the special orthogonal group $SO(3)$. In *Proc. of the* 44*th IEEE Conf. on Decision and Control*, pages 1477–1484.

Martin, P. and Salaün, E. (2007). Invariant observers for attitude and heading estimation from low-cost inertial and magnetic sensors. In *Proc. of the* 46*th IEEE Conf. on Decision and Control*.

Stevens, B. and Lewis, F. (2003). *Aircraft control and simulation*. Wiley, second edition.

Thienel, J. and Sanner, R. (2003). A coupled nonlinear spacecraft attitude controller and observer with an unknown constant gyro bias and gyro noise. *IEEE Trans. Automat. Control*, 48(11):2011–2015.

## APPENDIX: QUATERNIONS

Thanks to their four coordinates, quaternions provide a global parametrization of the orientation of a rigid body (whereas a parametrization with three Euler angles necessarily has singularities). Indeed, to any quaternion $q$ with unit norm is associated a rotation matrix $R_q \in SO(3)$ by

$$q^{-1} * \vec{p} * q = R_q \cdot \vec{p} \quad \text{for all } \vec{p} \in \mathbb{R}^3.$$

A quaternion $p$ can be thought of as a scalar $p_0 \in \mathbb{R}$ together with a vector $\vec{p} \in \mathbb{R}^3$,

$$p = \begin{pmatrix} p_0 \\ \vec{p} \end{pmatrix}.$$

The (non commutative) quaternion product $*$ then reads

$$p * q \triangleq \begin{pmatrix} p_0 q_0 - \vec{p} \cdot \vec{q} \\ p_0 \vec{q} + q_0 \vec{p} + \vec{p} \times \vec{q} \end{pmatrix}.$$

The unit element is $e \triangleq \begin{pmatrix} 1 \\ \vec{0} \end{pmatrix}$, and $(p * q)^{-1} = q^{-1} * p^{-1}$.

Any scalar $p_0 \in \mathbb{R}$ can be seen as the quaternion $\begin{pmatrix} p_0 \\ \vec{0} \end{pmatrix}$, and any vector $\vec{p} \in \mathbb{R}^3$ can be seen as the quaternion $\begin{pmatrix} 0 \\ \vec{p} \end{pmatrix}$. We systematically use these identifications in the paper, which greatly simplifies the notations.

We have the useful formulas

$$p \times q \triangleq \vec{p} \times \vec{q} = \frac{1}{2}(p * q - q * p)$$

$$(\vec{p} \cdot \vec{q})\vec{r} = -\frac{1}{2}(p * q + q * p) * r.$$

If $q$ depends on time, then $\dot{q}^{-1} = -q^{-1} * \dot{q} * q^{-1}$.

Finally, consider the differential equation $\dot{q} = q * u + v * q$ where $u, v$ are vectors $\in \mathbb{R}^3$. Let $q^T$ be defined by $\begin{pmatrix} q_0 \\ -\vec{q} \end{pmatrix}$. Then $q * q^T = \|q\|^2$. Therefore,

$$\dot{\overbrace{q * q^T}} = q * (u + u^T) * q^T + \|q\|^2 (v + v^T) = 0$$

since $u, v$ are vectors. Hence the norm of $q$ is constant.

# SHORT PAPERS

# RECURSIVE AND BACKWARD REASONING IN THE VERIFICATION ON HYBRID SYSTEMS

Stefan Ratschan
*Institute of Computer Science, Czech Academy of Sciences, Prague, Czech Republic*
*stefan.ratschan@cs.cas.cz*

Zhikun She
*LMIB and School of Science, Beihang University, Beijing, China*
*zhikun.she@buaa.edu.cn*

Keywords: Hybrid Systems, Verification, Constraint Propagation.

Abstract: In this paper we introduce two improvements to the method of verification of hybrid systems by constraint propagation based abstraction refinement that we introduced earlier. The first improvement improves the recursive propagation of reachability information over the regions constituting the abstraction, and the second improvement reasons backward from the set of unsafe states, instead of reasoning forward from the set of initial states. Detailed computational experiments document the usefulness of these improvements.

## 1 INTRODUCTION

Safety verification of hybrid systems is the problem of verifying that for a given hybrid system no trajectory that starts in an initial state ever reaches an unsafe state. Abstraction refinement approaches this problem by iteratively refining an overapproximation of the hybrid system (the *abstraction*) that is constructed in such a way that the safety of the abstraction implies the safety of the concrete system. In our method of constraint propagation based abstraction refinement (Ratschan and She, 2007) the abstraction is built by decomposing the state-space into hyper-rectangles (*boxes*) and using a constraint solver to test, which of these boxes might contain an initial/unsafe state, and which box might be reachable from another box.

In this paper, we introduce two improvements to the method: recursive reasoning and backward reasoning. Recursive reasoning improves the way the method removes elements from boxes for which it can prove that they are not reachable from an initial state. The original method argues that a point in a box is not reachable from another box, if it is not reachable from a point on the common boundary. In this paper we strengthen this condition using a convenient over-approximation of the requirement that this common point on the boundary again has to be reachable. Backward reasoning uses the observation that we can remove not only elements from boxes for which we can prove that they are not reachable from an initial state, but also elements for which we can prove that they do not lead to an unsafe state.

There are various other methods for the verification of hybrid systems that use a decomposition of the state space into boxes (Preußig et al., 1999; Kloetzer and Belta, 2006). Another paper (Frehse et al., 2006) employs backward reasoning in a more coarse-grained manner than in this paper, computing over-approximations of increasingly precise forward and backward reach sets.

The content of the paper is as follows: In Section 2 we review our hybrid systems formalism, and in Section 3 we review our verification method and discuss properties of the underlying constraint solving technique; in Sections 4 and 5 we introduce the first improvement to our verification method, and in Section 6 our second improvement and the combination of the two improvements; in Section 7 we present some computational experiments, and in Section 8 we conclude the paper.

## 2 VERIFICATION OF HYBRID SYSTEMS

Hybrid systems are systems with continuous and discrete state variables. In this paper, we briefly recall our formalism for modeling hybrid systems (Ratschan and She, 2007).

We use a set $S$ to denote the modes of a hybrid

system, where $S$ is finite and nonempty. $I_1,\ldots,I_k \subseteq \mathbb{R}$ are compact intervals over which the continuous variables of a hybrid system range. $\Phi$ denotes the state space of a hybrid system, i.e., $\Phi = S \times I_1 \times \cdots \times I_k$.

**Definition 1.** *A* hybrid system *$H$ is a tuple $(Flow, Jump, Init, Unsafe)$, where $Flow \subseteq \Phi \times \mathbb{R}^k$, $Jump \subseteq \Phi \times \Phi$, $I \subseteq \Phi$, and $Unsafe \subseteq \Phi$.*

Informally speaking, the predicate *Init* specifies the initial states of a hybrid system and *Unsafe* the states that should not be reachable from an initial state. The relation *Flow* specifies how the system may develop continuously by relating each state to the possible corresponding derivatives, and *Jump* specifies how $H$ may change states discontinuously by relating each state to its possible successor states. Formally, the behavior of $H$ is defined as follows:

**Definition 2.** *A* flow *of length $l \geq 0$ in a mode $s \in S$ is a function $r : [0,l] \to \Phi$ such that the projection of $r$ to its continuous part is differentiable and for all $t \in [0,l]$, the mode of $r(t)$ is $s$. A* trajectory *of $H$ is a sequence of flows $r_0,\ldots,r_p$ of lengths $l_0,\ldots,l_p$ such that for all $i \in \{0,\ldots,p\}$,*

1. *if $i > 0$ then $(r_{i-1}(l_{i-1}), r_i(0)) \in Jump$, and*
2. *if $l_i > 0$ then $(r_i(t), \dot{r}_i(t)) \in Flow$, for all $t \in [0,l_i]$, where $\dot{r}_i$ is the derivative of the projection of $r_i$ to its continuous component.*

**Definition 3.** *A (concrete)* counterexample *of a hybrid system $H$ is a trajectory $r_0,\ldots,r_p$ of $H$ such that $r_0(0) \in Init$ and $r_p(l) \in Unsafe$, where $l$ is the length of $r_p$. $H$ is* safe *if it does not have a counterexample.*

We use the following constraint language to describe hybrid systems and corresponding safety verification problems. The variable $s$ ranges over $S$ and the tuple of variables $\vec{x} = (x_1,\ldots,x_k)$ ranges over $I_1 \times \cdots \times I_k$, respectively. In addition, to denote the derivatives of $x_1,\ldots,x_k$ we use the tuple of variables $\dot{\vec{x}} = (\dot{x}_1,\ldots,\dot{x}_k)$ that ranges over $\mathbb{R}^k$, and to denote the targets of jumps, we use the primed variable $s'$ and the tuple of variables $\vec{x}' = (x'_1,\ldots,x'_k)$ that range over $S$ and $I_1 \times \cdots \times I_k$, respectively. Constraints are arbitrary Boolean combinations of equalities and inequalities over terms that may contain function symbols, such as $+$, $\times$, exp, sin, and cos.

We assume in the remainder of the text that a hybrid system is described by our constraint language. That means, the flows of a hybrid system are given by a constraint $Flow(s,\vec{x},\dot{\vec{x}})$, the jumps are given by a constraint $Jump(s,\vec{x},s',\vec{x}')$, the initial states are given by a constraint $Init(s,\vec{x})$, and a constraint $Unsafe(s,\vec{x})$ describes the unsafe states. To simplify notation, we do not distinguish between a constraint and the set it represents.

**Example 1.** *Consider the following simple hybrid system with the modes $m_1, m_2$ and the continuous variables $x_1, x_2$ which both range over the interval $[0,2]$, i.e, $\Phi = \{m_1,m_2\} \times [0,2] \times [0,2]$.*

*The set of initial states are given by the $Init(s,(x_1,x_2)) = (s = m_1 \wedge x_1 = 0 \wedge x_2 = 0)$. The constraint $Unsafe(s,(x_1,x_2)) = (x_1 > 1.5 \wedge x_2 = 1.5)$ describes the set of unsafe states. The hybrid system can switch modes from $m_1$ to $m_2$ if $x_2 = 1$, i.e., $Jump(s,(x_1,x_2),s',(x'_1,x'_2)) = (s = m_1 \wedge x_2 = 1) \to (s' = m_2 \wedge x'_1 = x_1 \wedge x'_2 = x_2)$ . The continuous behavior is described by constants. In addition, for a flow in mode $m_1$, the constraint $0 \leq x_1 \leq 1$ must hold. The corresponding flow constraint is*

$$\begin{gathered} Flow(s,(x_1,x_2),(\dot{x}_1,\dot{x}_2)) = \\ (s = m_1 \to (\dot{x}_1 = 1 \wedge \dot{x}_2 = 1 \wedge 0 \leq x_1 \leq 1)) \wedge \\ (s = m_2 \to (\dot{x}_1 = 1 \wedge \dot{x}_2 = -1)) \,. \end{gathered}$$

*Note that the constraint $0 \leq x_1 \leq 1$ in flow forces a jump from mode $m_1$ to $m_2$ if $x_1$ becomes 1.*

*Obviously, this hybrid system is safe.* ■

# 3 FORWARD SEARCH BASED ABSTRACTION REFINEMENT

In this section, we review our previous approach (Ratschan and She, 2007) for verifying safety of hybrid systems using constraint propagation based abstraction refinement.

We abstract to systems of the following form:

**Definition 4.** *A* discrete system *over a finite set $S$ is a tuple $(Trans, Init, Unsafe)$ where $Trans \subseteq S \times S$ and $Init \subseteq S$, $Unsafe \subseteq S$. We call the set $S$ the* state space *of the system.*

In contrast to Definition 1, here the state space is a parameter. This will allow us to add new states to the state space during abstraction refinement.

**Definition 5.** *A* trajectory of a discrete system *$(Trans, Init, Unsafe)$* over a set *$S$ is a function $r : \{0,\ldots,p\} \mapsto S$ such that for all $t \in \{1,\ldots,p\}$, $(r(t-1), r(t)) \in Trans$. The system is* safe *if and only if there is no trajectory from an element of Init, to an element of Unsafe.*

When we use abstraction to analyze hybrid systems, the abstraction should over-approximate the concrete system in a conservative way: if the abstraction is safe, then the original system should also be safe. If the current abstraction is not yet safe, we refine the abstraction, that is, we include more information about the concrete system into it. This results in Algorithm 1.

**Algorithm 1:** Abstraction Refinement.

**Require:** a hybrid system $H$ described by constraints
**Ensure:** "safe", if the algorithm terminates
  let $A$ be a discrete abstraction of the hybrid system represented by $H$
  **while** $A$ is not safe **do**
    refine the abstraction $A$
  **end while**

In order to implement this algorithm, we need to fix the state space of the abstract system. Here we use pairs $(s,B)$, where $s$ is one of the modes $\{\mathtt{s_1},\ldots,\mathtt{s_n}\}$ and $B$ is a hyper-rectangle (*box*), representing subsets of the concrete state space $\Phi$. Together with an abstract state, we store the information whether it is initial or unsafe and the information from which other states it is reachable. We call such information the *marks* of the state. For the initial abstraction we use the state space $\{(\mathtt{s_i},\{\vec{x} \mid (\mathtt{s_i},\vec{x}) \in \Phi\}) \mid 1 \leq i \leq n\}$, where all states are marked as initial, and unsafe, and all transitions between states are possible.

For refining the abstraction, we split a box into two pieces, replace one abstract state by two, and include more information from the concrete system into the abstract one by removing unreachable elements from the boxes, removing superfluous marks from the new abstract states, and removing unreachable states from the abstraction.

To remove unreachable elements from the boxes representing the abstraction, we use a constraint that formalizes when an element of the concrete state space might be reachable, and then remove elements that do not fulfill this constraint. In order to do this, for a box $B = [\underline{x}_1,\overline{x}_1] \times \cdots \times [\underline{x}_k,\overline{x}_k]$, we let its $j$-th lower face be $[\underline{x}_1,\overline{x}_1] \times \cdots \times [\underline{x}_j,\underline{x}_j] \times \cdots \times [\underline{x}_k,\overline{x}_k]$ and its $j$-th upper face be $[\underline{x}_1,\overline{x}_1] \times \cdots \times [\overline{x}_j,\overline{x}_j] \times \cdots \times [\underline{x}_k,\overline{x}_k]$. Two boxes are *non-overlapping* if their interiors are disjoint.

Now observe that a point in a box $B$ is reachable only if it is reachable either from the initial set via a flow in $B$, from a jump via a flow in $B$, or from a neighboring box via a flow in $B$. We will now formulate constraints corresponding to each of these conditions. Then we can remove points from boxes that do not fulfill at least one of these constraints.

The approach can be used with any constraint that describes that $\vec{y}$ can be reachable from $\vec{x}$ via a flow in $B$ and mode $s$, for example, the one introduced in our previous publications (Ratschan and She, 2006). We denote the used constraint by $Reach_B(s,\vec{x},\vec{y})$. Thus, the above three possibilities for reachability allow us to formulate the following theorem:

**Theorem 1.** *For a set of abstract states $\mathcal{B}$, a pair $(s',B') \in \mathcal{B}$ and a point $\vec{z} \in B'$, if $(s',\vec{z})$ is reachable and z is not an element of the box of any other abstract state in $\mathcal{B}$, then*

$$Ifl_{B'}(s',\vec{z}) \vee \bigvee_{(s,B)\in\mathcal{B}} Jfl_{B,B'}(s,s',\vec{z}) \vee \bigvee_{(s,B)\in\mathcal{B},s=s',B\neq B'} Bfl_{B,B'}(s',\vec{z})$$

*where $Ifl_{B'}(s',\vec{z})$, $Jfl_{B,B'}(s,s',\vec{z})$, and $Bfl_{B,B'}(s',\vec{z})$ denote the following three constraints, respectively:*

- $\exists \vec{x} \in B'\,[Init(s',\vec{x}) \wedge Reach_{B'}(s',\vec{x},\vec{z})]$,
- $\exists \vec{x} \in B \exists \vec{x}' \in B'[Jump(s,\vec{x},s',\vec{x}') \wedge Reach_{B'}(s',\vec{x}',\vec{z})]$
- $\exists \vec{x} \in B \cap B'[[\forall faces\ F\ of\ B'[\vec{x} \in F \Rightarrow in^F_{s',B'}(\vec{x})]] \wedge Reach_{B'}(s',\vec{x},\vec{z})]$.

*Here, $in^F_{s',B'}(\vec{x}) = \exists \dot{x}_1,\ldots,\exists \dot{x}_k[F(s',\vec{x},(\dot{x}_1,\ldots,\dot{x}_k)) \wedge \dot{x}_j \geq 0]$, if $F$ is the j-th lower face of $B'$, and if $F$ is the j-th upper face of $B'$, $incoming^F_{s',B'}(\vec{x}) = \exists \dot{x}_1,\ldots,\exists \dot{x}_k[F(s',\vec{x},(\dot{x}_1,\ldots,\dot{x}_k)) \wedge \dot{x}_j \leq 0]$.*

We denote the main constraint of Theorem 1 by $reach_{\mathcal{B},B'}(s',\vec{z})$. If we can prove that a certain point does not fulfill this constraint, we know that it is not reachable. For now, we assume that we have an algorithm (a *pruning algorithm*) that takes such a constraint, and an abstract state $(s',B')$ and returns a subbox of $B'$ that still contains all the solutions of the constraint in $B'$. Since the constraint $reach_{\mathcal{B},B'}(s',\vec{z})$ depends on all current abstract states, a change of $B'$ might allow further pruning of other abstract states. So we can repeat pruning until a fixpoint is reached. Given a set of abstract states $\mathcal{B}$, we denote the resulting fixpoint by $Prune_H(\mathcal{B})$.

Now we remove the initial mark from an abstract state $(s',B')$ if we can disprove $Ifl_{B'}(s',\vec{z}')$ in Theorem 1 (i.e., if the pruning algorithm returned the empty box for this constraint), and we remove the unsafe mark of an abstract state state $(s',B')$ if we can disprove the constraint $\exists \vec{x} \in B\ Unsafe(s,\vec{x})$. Moreover, we remove a transition from $(s,B)$ to $(s',B')$ if we can disprove both $Bfl_{B,B'}(s',\vec{z}')$ and $Jfl_{B,B'}(s,s',\vec{z}')$ from Theorem 1. As already mentioned, after recomputing the marks, we remove all abstract states from the abstraction that are not reachable. It is easy to compute these, since the set of abstract states is finite.

There are several methods for implementing the needed pruning algorithms (Benhamou and Granvilliers, 2006). For the domain of the real numbers, given a constraint $c$ and a floating-point box $B$, they compute another floating-point box $P(c,B)$ such that $P(c,B) \subseteq B$ (contractance), and such that $P(c,B)$ contains all solutions of $c$ in $B$. Existential quantifiers and disjunctions can be handled by slight extensions (for disjunctions we take the *box union* $\uplus$).

Such pruning algorithms $P$ usually have the *monotonicity property* that for a constraint $c$, and boxes $B$ and $B'$ with $B' \subseteq B$, $P(c,B') \subseteq P(c,B)$. Moreover, in practice, if $B' \subseteq B$ then $P(c,B')$ is often much smaller than $P(c,B)$. We will exploit this in the improvement of our method described in the next section. In addition, it pays off to distribute disjunctions over conjunctions:

**Lemma 1.** *For constraints $c_1,\ldots,c_n,d$ and a box $B$,*

$$P(\vee_{i\in\{1,\ldots,n\}}(c_i \wedge d),B) \subseteq P((\vee_{i\in\{1,\ldots,n\}}c_i)\wedge d,B)$$

**Proof.** For each $i \in \{1,\ldots,n\}$, $P(c_i \wedge d,B) \subseteq P((\vee_{i\in\{1,\ldots,n\}}c_i) \wedge d,B)$. Thus, $\uplus_{i\in\{1,\ldots,n\}}P(c_i \wedge d,B) \subseteq P((\vee_{i\in\{1,\ldots,n\}}c_i) \wedge d,B)$. Since $P(\vee_{i\in\{1,\ldots,n\}}(c_i\wedge d),B) = \uplus_{i\in\{1,\ldots,n\}}P(c_i\wedge d,B)$, the lemma holds. ■

# 4 RECURSIVE PRUNING

In this section we introduce the first improvement to the verification method described in Section 3. Throughout the rest of the paper we assume an abstraction consisting of a set of abstract states $\mathcal{B}$. The improvement introduced in this section aims at pruning more unreachable states from $\mathcal{B}$ by improving the recursive propagation of reachability information for flows from one box to the next.

We consider the pruning of an abstract state $(s',B') \in \mathcal{B}$. The constraint $Bfl_{B,B'}(s',\vec{z})$ defined within Theorem 1 models the fact that a certain point $\vec{z}$ in the box $B'$ is reached from a neighboring box $B$ of $B'$ via a flow in $B'$. This flow reaches $\vec{z}$ through a common point $\vec{x} \in B\cap B'$ (see Figure 1).

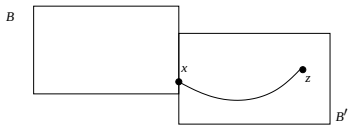



Figure 1: Recursive Pruning.

The basic idea upon which we build in this section is to strengthen this constraint by requiring that also $\vec{x}$ be reachable in the neighboring box $B$. Naively, this could be done by adding the constraint $reach_{\mathcal{B},B}(s',\vec{x})$ to the constraint $Bfl_{B,B'}(s',\vec{z})$. However, since $Bfl_{B,B'}(s',\vec{z})$ is itself a part of $reach_{\mathcal{B},B'}(s',\vec{x})$, this would result in an infinitely large constraint due to recursion. One could make the constraint finite, by bounding the recursion, but this still would result in a very large constraint. We avoid this, by observing that the neighboring box $B$ is already the result of pruning wrt. $reach_{\mathcal{B},B}(s',\vec{x})$. However, a part of this information is lost because we first prune $B$ and only then take the intersection $B\cap B'$ (i.e., we compute $P(reach_{\mathcal{B},B}(s',\vec{x}),B)\cap B'$), and we have:

**Lemma 2.**

$$P(reach_{\mathcal{B},B}(s',\vec{x}),B\cap B') \subseteq P(reach_{\mathcal{B},B}(s',\vec{x}),B)\cap B'$$

**Proof.** Due to monotonicity of constraint propagation, $P(reach_{\mathcal{B},B}(s',\vec{x}),B \cap B')$ is a subset of $P(reach_{\mathcal{B},B}(s',\vec{x}),B)$. Moreover, $P(reach_{\mathcal{B},B}(s',\vec{x}),B \cap B') \subseteq P(reach_{\mathcal{B},B}(s',\vec{x}),B')$, and hence $P(reach_{\mathcal{B},B}(s',\vec{x}),B \cap B') \subseteq B'$. So $P(reach_{\mathcal{B},B}(s',\vec{x}),B \cap B')$ is also a subset of the intersection of $P(reach_{\mathcal{B},B}(s',\vec{x}),B)$ and $B'$. ■

In practice, the set on the left-hand side might be significantly smaller than the set on the right-hand side (i.e., than the set currently used in the method in Section 3). So it makes sense to compute $P(reach_{\mathcal{B},B'}(s',\vec{x}),B \cap B')$ instead of $P(reach_{\mathcal{B},B}(s',\vec{x}),B)\cap B'$. This means that in addition to pruning each box in the abstraction, we could also prune the intersection between each pair of boxes. However, this would need a quadratical number of prunings and stored boxes in memory.

To avoid this, we use an over-approximation of $P(reach_{\mathcal{B},B}(s',\vec{x}),B\cap B')$ that is still a subset of $P(reach_{\mathcal{B},B}(s',\vec{x}),B)\cap B'$. We use the information that the boxes of our abstraction are non-overlapping (i.e., even if two boxes intersect, they only share the boundary but no points of the interior). This implies that the intersection $B\cap B'$ will always be a subset of the boundary of $B$—independent of the form of the box $B'$. So one could try to use the boundary $\square B$ of $B$ instead of the box $B\cap B'$ when computing $P(reach_{\mathcal{B},B}(s',\vec{x}),B\cap B')$. However, since $\square B$ is not a box and hence it cannot be an argument to the pruning function, we apply the pruning function to its constituent faces separately. That is, we use the constraint that expresses a disjunction over all faces:

$$\bigvee_{F,\text{face of } B} \left[\vec{x} \in F \wedge reach_{\mathcal{B},B}(s',\vec{x})\right]$$

and call this constraint $reachbound_{\mathcal{B},B}(s',\vec{x})$. Although this over-approximates $P(reach_{\mathcal{B},B}(s',\vec{x}),B\cap B')$, Lemma 2 still holds in analogy:

**Lemma 3.** $P(\bigvee_{F,face\ of\ B}\left[\vec{x}\in F\wedge reach_{\mathcal{B},B}(s',\vec{x})\right],B) \cap B' \subseteq P(reach_{\mathcal{B},B}(s',\vec{x}),B)\cap B'$.

**Proof.** The disjunction is pruned by taking the box union over the result of pruning each disjunct. Since each face of $B$ is a subset of $B$, due to monotonicity of constraint propagation, for each face $F$, $P(reach_{\mathcal{B},B}(s',\vec{x}),F) \subseteq P(reach_{\mathcal{B},B}(s',\vec{x}),B)$. Hence

also the box union over the result of pruning each disjunct is a subset of $P(reach_{\mathcal{B},B}(s',\vec{x}),B)$, which implies the lemma. ■

Since the constraint on the left-hand side only depends on one box, we can compute the corresponding pruning $P(reachbound_{\mathcal{B},B}(s',\vec{x}),B)$ only for one abstract state, and store the resulting box with that abstract state. Since this box encloses the set of states where a flow might leave the abstract state, we call it the *outflow-box* of the abstract state. So, instead of $B \cap B'$ in the constraint $Bfl_{B,B'}$ we can now take the outflow-box of $B$, and due to Lemma 3 we will arrive at a result that is at least as tight as before.

This is illustrated on an example in Figure 2, where the dotted box is the *outflow-box* resulting from a situation where the upper and left face of box $B$ have been pruned to the empty set, and the outflow-box is the result of taking the union of the result of pruning the two other faces.

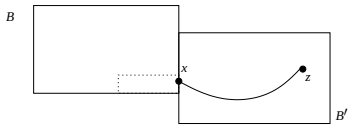



Figure 2: Pruning Faces.

Note that splitting a box $B$ representing a certain abstract state changes its faces. Especially, there might be trajectories that leave the resulting boxes through the new face along which $B$ has been split. Hence the outflow-box of this abstract state becomes invalid. So we simply set the outflow-box to the whole box $B$ and re-compute it, the next time $B$ is pruned.

# 5 RECURSIVE PRUNING WITH OUTGOING CONDITION

In the previous section we used the fact that within the constraint $Bfl$ we can exploit the information that the common point $\vec{x} \in B \cap B'$ itself has to be reachable. In this section we strengthen this information by observing that in order for a trajectory to be able to leave the box $B$ to enter the box $B'$, the vector field at $x$ has to point out of $B$.

This can be modelled by adding an additional condition in the constraint $reachbound_{\mathcal{B},B}(s',\vec{x})$, arriving at

$$\bigvee_{F,\text{face of } B} \left[\vec{x} \in F \wedge reach_{\mathcal{B},B}(s',\vec{x}) \wedge out^F_{s',B}(\vec{x})\right],$$

where $out^F_{s',B}(\vec{x})$ is equal to $in^F_{s',B}(\vec{x})$ with the inequality sign switched. Now, since $reach_{\mathcal{B},B}(s',\vec{x})$ is a disjunction, Lemma 1 suggests to improve it by pulling out the new conjunction of $reachbound_{\mathcal{B},B}(s',\vec{x})$, arriving at

$$\begin{aligned}
&\bigvee_{F,\text{face of}B} \left[\vec{x} \in F \wedge out^F_{s',B}(\vec{x}) \wedge Ifl_B(s',\vec{x})\right] \vee \\
&\bigvee_{(s,B')\in\mathcal{B}} \Big[ \bigvee_{F,\text{face of}B} [\vec{x} \in F \wedge out^F_{s',B}(\vec{x}) \\
&\qquad\qquad \wedge Jfl_{B',B}(s,s',\vec{x})]\Big] \vee \\
&\bigvee_{\substack{(s,B')\in\mathcal{B} \\ s=s',B'\neq B}} \Big[ \bigvee_{F,\text{face of}B} [\vec{x} \in F \wedge out^F_{s',B}(\vec{x}) \\
&\qquad\qquad \wedge Bfl_{B',B}(s',\vec{x})]\Big].
\end{aligned}$$

We call the resulting constraint $reachout_{\mathcal{B},B}(s',\vec{x})$, and use this constraint instead to compute the outflow-box of each abstract state.

The following examples illustrates the improvement provided by *reachout* over *reachbound*: Consider the differential equation $(\dot{x}_1,\dot{x}_2) = (1,1)$ with a box $B = [0,1]\times[0,1]$ and an initial point $x_0 = (0,0)$. If we prune a face $[1,1]\times[0,1]$ or $[0,1]\times[1,1]$ wrt. *reachbound*, we will get the point $(1,1)$; and if we prune a face $[0,0]\times[0,1]$ or $[0,1]\times[0,0]$ wrt. *reachbound*, we will get the point $(0,0)$. That is, if we apply the pruning algorithm to *reachbound* and $B$, we will get the full box $[0,1]\times[0,1]$. Only when adding the outgoing condition, arriving at the constraint *reachout*, we can ignore trajectories moving into the box, arriving at the point $[1,1]\times[1,1]$.

# 6 BACKWARD REASONING AND COMBINATION

As described in Section 3, in our method we remove elements from the state space for which we can prove that they are not reachable from an initial state. However, the task of safety verification is to prove the absence of a trajectory that starts in an initial state *and reaches an unsafe state*. Hence we can also remove elements from the state space for which we can prove that they do not lead to an unsafe state—without destroying the property that safety of the abstraction implies safety of the concrete system.

For this, observe that a point might lead to an unsafe state only if there is a flow from this point to the unsafe set directly, or a flow from this point to a jump, or a flow from this point to a boundary point. Hence we can formulate an analogous version of Theorem 1:

**Theorem 2.** *For a set of abstract states $\mathcal{B}$, a pair $(s',B') \in \mathcal{B}$ and a point $\vec{z} \in B'$, if the unsafe set is reachable from $(s',\vec{z})$ and $\vec{z}$ is not an element of the box of any other abstract state in $\mathcal{B}$, then*

$$Urev_{B'}(s',\vec{z}) \vee \bigvee_{(s,B)\in\mathcal{B}} Jrev_{B,B'}(s,s',\vec{z}) \vee \bigvee_{(s,B)\in\mathcal{B},s=s',B\neq B'} Brev_{B,B'}(s',\vec{z}),$$

*where $Urev_{B'}(s',\vec{z})$, $Jrev_{B,B'}(s,s',\vec{z})$, and $Brev_{B,B'}(s',\vec{z})$ denote the following three constraints, respectively:*

- $\exists \vec{x} \in B[Reach_B(s,\vec{z},\vec{x}) \wedge Unsafe(s,\vec{x})]$,
- $\exists \vec{x} \in B \exists \vec{x}' \in B'[Reach_B(s,\vec{z},\vec{x}) \wedge Jump(s,\vec{x},s',\vec{x}')]$
- $\exists \vec{x} \in B \cap B'[Reach_B(s,\vec{z},\vec{x}) \wedge [\forall$ *faces* $F$ *of* $B[\vec{x} \in F \Rightarrow in^F_{s,B'}(x)]]]$

In a similar way as forward reasoning, backward reasoning also allows us to update the initial/unsafe marks and transitions of the abstraction.

Note that by using forward and backward reasoning in Algorithm 1 we might succeed in removing *all* elements from the concrete state space. This results in an empty abstraction which is trivially safe. Hence, Algorithm 1 can report a successful verification in this case. However, the combination of recursive pruning with backward pruning introduces additional difficulties: the outflow box is computed using forward reasoning, and when a box is changed due to backward reasoning, its outflow box is not valid any more. We solve this problem by always, first applying forward pruning and then backward pruning. If backward pruning changes the box, we apply forward pruning again which recomputes a valid outflow box.

## 7 EXPERIMENTAL RESULTS

We extended our hybrid systems verification package HSOLVER (Ratschan and She, 2004) with the two improvements introduced in this paper. Then we used our problem database[1] of hybrid systems to evaluate our improvements. The experimental results are summarized in Table 1, Table 2 and Table 3 for different versions.

We used an IBM notebook with an Intel Pentium 1.70 GHz CPU with 1024 Mbytes of main memory running Linux. The running times are in seconds and the computations were cancelled when computation did not terminate before three hours or the number of the abstract states exceeded 1000. We used the default splitting strategy of HSOLVER.

[1] `http://hsolver.sourceforge.net/benchmarks`

Table 1: Experimental Results: I.

| | Forward | |
|---|---|---|
| Example | time | splits |
| 1-flow | unknown | |
| 2-tanks | 1.12 | 31 |
| car | 0.47 | 0 |
| circuit | 53.80 | 186 |
| clock | 1.99 | 32 |
| convoi-1 | 1.06 | 0 |
| convoi | 2157.26 | 374 |
| eco | 125.24 | 223 |
| focus | 3.59 | 57 |
| mixing | 296.67 | 174 |
| mutant | 7286.40 | 742 |
| real-eigen | 0.59 | 2 |
| s-focus | 0.54 | 2 |
| trivial-hard | 0.76 | 26 |
| van-der-pole (VDP) | 25.42 | 64 |

Table 2: Experimental Results: II.

| | Backward | | For-Backward | |
|---|---|---|---|---|
| Example | time | splits | time | splits |
| 1-flow | unknown | | 0.32 | 1 |
| 2-tanks | 0.10 | 6 | 0.43 | 4 |
| car | unknown | | 1.05 | 0 |
| circuit | unknown | | 62.99 | 188 |
| clock | 35.79 | 327 | 2.00 | 43 |
| convoi-1 | unknown | | 1.65 | 0 |
| convoi | unknown | | 1847.67 | 300 |
| eco | unknown | | 18.32 | 52 |
| focus | 0.62 | 34 | 0.89 | 15 |
| mixing | 1.47 | 7 | 1.74 | 0 |
| mutant | unknown | | 8493.97 | 618 |
| real-eigen | unknown | | 0.61 | 0 |
| s-focus | unknown | | 0.44 | 1 |
| trivial-hard | 0.03 | 0 | 0.05 | 0 |
| VDP | 0.47 | 1 | 0.77 | 1 |

Comparing the forward version and backward version, there is no clear winner, although the forward version is successful in more cases. The reason seems to lie in the fact that for more examples the set of initial states is smaller than the set of unsafe states.

Moreover, the experimental results show that: (1) For most of the examples, the combined forward and backward version use less splitting steps than both the forward version and backward version. However, for some examples, the CPU time is worse since in the combined version, the constraints are more complex. Note that for some examples (e.g., circuit and clock), the combined version needs slightly more

Table 3: Experimental Results: III.

| | Recursive | | Rec-Backward | |
|---|---|---|---|---|
| Example | time | splits | time | splits |
| 1-flow | unknown | | 0.32 | 1 |
| 2-tanks | 0.26 | 3 | 0.28 | 1 |
| car | 1.29 | 0 | 1.07 | 0 |
| circuit | 53.12 | 171 | 68.28 | 192 |
| clock | 1.97 | 16 | 0.74 | 14 |
| convoi-1 | 2.70 | 0 | 1.71 | 0 |
| convoi | 2177.75 | 374 | 1830.56 | 300 |
| eco | 253.19 | 290 | 5.77 | 22 |
| focus | 2.79 | 48 | 0.42 | 8 |
| mixing | 104.96 | 109 | 1.75 | 0 |
| mutant | 1978.43 | 195 | 1850.17 | 191 |
| real-eigen | 0.59 | 2 | 0.61 | 0 |
| s-focus | 0.53 | 2 | 0.44 | 1 |
| trivial-hard | 0.07 | 4 | 0.05 | 0 |
| VDP | 25.81 | 64 | 0.77 | 1 |

splitting steps. The reason is that although the combined version is more successful in pruning, the splitting heuristics will sometimes choose different boxes which then results—in rare cases—in more necessary splits. (2) For all examples except one, the recursive version needs less splitting steps than the forward version. However, for the eco example, the recursive version needs more splitting steps. This is due to the same reason as above—more successful pruning leads to different box choices. Again, for some example the CPU time is worse since the constraints in the recursive version are more complex. (3) Again with one exception (circuit), the combined recursive and backward version always needs less splitting steps than both the recursive version and combined forward and backward version, often even much less. For most examples also the run-time improved, sometimes over an order of magnitute. Only for two additional, rather easy examples (car, convoi-1), the CPU time slightly increases since the constraints in the combined recursive and backward version are more complex.

Summarizing, the contributions of this paper result in a definite and robust efficiency improvement of the algorithms.

## 8 CONCLUSIONS

In this paper we have introduced two improvements to a method of safety verification of hybrid systems by constraint propagation based abstraction refinement. The provided computational experiments clearly show the advantage of proposed improvements. We will base future improvements of the method on a detailed study of the behavior of the used algorithms on further benchmark problems.

## ACKNOWLEDGEMENTS

The work of the first author has been supported by GAČR grant 201/08/J020 and by the institutional research plan AV0Z100300504. The second author was partly supported by the National Key Basic Research Program of China under Grant No. 2005CB321902 and the Program for Excellent Talents of Beijing under Grant No. 20071D1600600410.

## REFERENCES

Benhamou, F. and Granvilliers, L. (2006). Continuous and interval constraints. In Rossi, F., van Beek, P., and Walsh, T., editors, *Handbook of Constraint Programming*, pages 571–603. Elsevier Amsterdam.

Frehse, G., Krogh, B. H., and Rutenbar, R. A. (2006). Verifying analog oscillator circuits using forward/backward abstraction refinement. In *DATE 2006: Design, Automation and Test in Europe*.

Kloetzer, M. and Belta, C. (2006). Reachability analysis of multi-affine systems. In Hespanha, J. and Tiwari, A., editors, *HSCC'06*, volume 3927 of *LNCS*. Springer.

Preußig, J., Stursberg, O., and Kowalewski, S. (1999). Reachability analysis of a class of switched continuous systems by integrating rectangular approximation and rectangular analysis. In Vaandrager, F. and van Schuppen, J., editors, *HSCC'99*, number 1569 in LNCS. Springer.

Ratschan, S. and She, Z. (2004). HSOLVER. http://hsolver.sourceforge.net. Software package.

Ratschan, S. and She, Z. (2006). Constraints for continuous reachability in the verification of hybrid systems. In *Proc. AISC'2006*, number 4120 in LNCS. Springer.

Ratschan, S. and She, Z. (2007). Safety verification of hybrid systems by constraint propagation based abstraction refinement. *ACM Transactions on Embedded Computing Systems*, 6(1).

# ON THE SAMPLING PERIOD IN STANDARD AND FUZZY CONTROL ALGORITHMS FOR SERVODRIVES

## *A Multicriterial Design and a Timing Strategy for Constant Sampling*

Dan Mihai
*University of Craiova, Decebal Blvd, 107, Craiova, Romania*
*dmihai@em.ucv.ro*

Keywords: Sampling period, PI algorithm, Fuzzy control, On-line timing, Servodrives.

Abstract: The paper deals with the best choice for the control sampling period in term of a multicriterial conditioning, with the on-line timing and with a comparison between the conventional (like PI) control algorithms and the fuzzy control. Several useful relations are followed by diagrams obtained in simulation and by different real-time recordings both for the timing and for characteristic variables of the system. Implementations with microcontroller and DSP are used for analyzing the design criterion and the timing strategy. The application field concerns a servodrive, so the real-time constraints are quite strong. The author conceived a general control strategy for the on-line timing based on imbricate interrupts, each pulse encoder acting on the hardware input interrupt of the control processor.

# 1 INTRODUCTION

The sampling frequency plays an essential role in implementing a digital control algorithm in real-time, especially for the fast systems. Most of the reference books give evaluations only for the upper limit of the sampling period (T), as if the ideal value would be as little as possible – only the capacity of the control processor being the constraint. The efficient choice of the sampling rate in closed-loop system is based on its influence on the performance of the control system. The absolute lower bound to the sample rate is set by the system bandwidth. The classical controllers (and their loops) are not robust and their tuning (including the additional T parameter) - although stated as well settled (Astrom, 1997), seems to be very difficult in complex conditions. Not a few applications and studies concern the sampling period design for different kind of fuzzy control. An earlier idea (Coleman, 1994) about the robustness comparison between fuzzy logic, PID control and sliding mode control, will be extended now in the area of the control sampling period.

During the last decades, the electrical drives field has integrated more and more design techniques, fast control processors and acquisition modules for high performance platforms. A hybrid approach is to have an inner current loop monitored by a fuzzy controller (Mrozek, 2000) while the main speed loop is monitored by a classical PI controller. Another approach is to design a self tuning fuzzy logic controller, based on some desired output behaviour and hence, does not requiring a precise model of the machine (Ibbini, 2002). Sometimes, the results are quite close in term of performance (in steady state or dynamic regime) but the implementation effort is much lower for the fuzzy solution (Silveira, 2002). Industrial equipment support from few kHz to 20 kHz sample rate for the velocity loop. This high rate of sampling combined with the velocity observer, allows equipment to provide a very fast control for industrial servo drives. Despite the availability of several high performance DSP controllers, many researchers are interested in designing optimal control algorithms based on low-cost solutions; such approach is typical for linear and sliding-mode controllers designed for a DC servo drive with a microcontroller and low sampling rate, typical for embedded systems (Kosek, 2007).

The author developed a general on-line timing for all digital control algorithms for the servodrives with a hardware position loop and a software one for the speed, proposing several relations for the multicriterial correlation of the involved parameters, T included. Several such relations require the T value to be superior to some threshold limits.

## 2 THE SERVODRIVE AND THE ON-LINE CONTROL TIMING

Figure 1a presents the main parts of the system. Each encoder pulse acts on the interrupt entry of the processor. The next control strategy and the timing remain the same for any motor (and its associated power supply), for different kind of control algorithms (standard, fuzzy). The general systemic structure is given by figure 1b; figure 1c is for a conventional control and the next one (1d) presents the fuzzy control. The main notations: $N_\alpha$*,$N_{\alpha k}$ - position set-point and the real position, in encoder pulses; $\varepsilon_{\alpha k}$, $\Delta\varepsilon_{\alpha k}$: position error, its variation referred to a sampling period; $\Delta N_{\alpha k}$ - pulses encoder during T; $c_k$, $c_{kout}$ - the computed control and its outputted value; $Norm_i$: normalization blocks for each Fuzzy Logic Controller (FLC); CPB: Control Processing Block; PS - Power supply with digital control input; $T_{gen}$ - torque generator; M - motor; En - encoder. The encode has $N_{p/r}$ pulses per revolution and the speed is monitored through a software image:

$$\omega_{ks} \approx \frac{\alpha_k - \alpha_{k-1}}{T} = \frac{2\pi \cdot k_{div} \cdot \Delta N_k}{N_{p/r} \cdot T} = c_{sp} \cdot \Delta N_k \qquad (1)$$

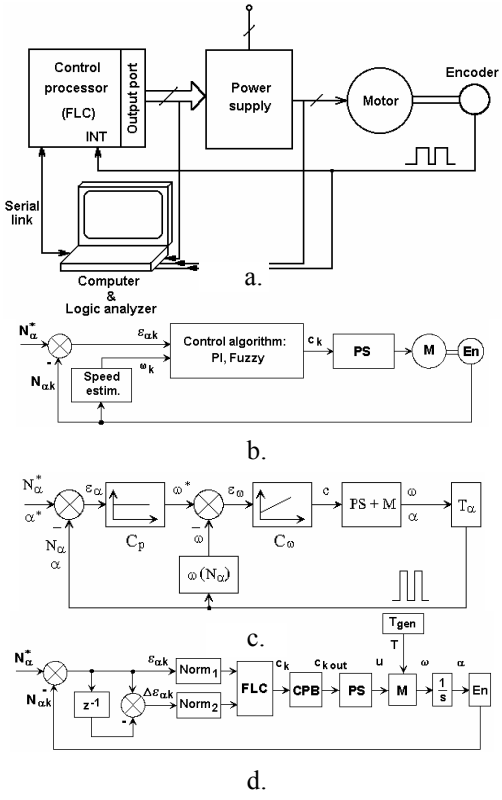



Figure 1: The drive with different controllers.

$k_{div}$ is a division / multiplication factor for encoder pulses. The real-time control strategy is based on 2 imbricate interrupts – figure 2. INT0 is a high level priority interrupt generated by encoder pulses (the falling edge). The low level priority interrupt is software generated, marking the sampling period - T. PULSE means the encoder signals. For standard or non-conventional control algorithms, T must be correlated with: a. the specific system dynamic; b. $n_{min}$ - the minimum accepted value of the real speed for which $\omega_{k\ soft}$ is detectable; c. $n_{lw}$ - the size of the data word (register) for $\omega_{ks}$; d. the accepted resolution for the speed; e. the program length of the on-line processing; f. The amount of memory available for on-line recordings. **b., c** and d**.** give the next restrictions for T:

$$\frac{60 \cdot k_{div}}{N_{p/r} \cdot n_{\min}\ [RPM]} \le T \le \frac{(2^{n_{lw}} - 1) \cdot 60 \cdot k_{div}}{N_{p/r} \cdot n_{\max}\ [RPM]} \qquad (2)$$

When the range speed covers a full data register, for having 1 LSB at minimum speed, T must be:

$$T \ge \frac{(2^{n_{reg}} - 1) \times 60 \times k_{div}}{N_{p/r} \times n_{\max}[RPM]} \qquad (3)$$

The e. condition generates another constraint for T (Mihai, 1999). INT0 requested by the falling edge encoder pulses computes $\varepsilon_{\alpha k}$, $\Delta N_k$ and needs the time-$\Delta t_{INT0}$. $\Delta t_{av.instr.}$ is the average time for a processor instruction. Having the code program length for the software interrupt routine-$N_{max.ALG.\ instr.}$ imposed by the algorithm, T must be:

$$T \ge \frac{N_{\max\ ALG\ instr.} \cdot \Delta t_{av.\ instr.}}{1 - \frac{n_{\max} \cdot N_{p/r} \cdot \Delta t_{INT0}}{60 \cdot k_{div}}} \qquad (4)$$

As for f. condition, the data memory space for all on-line records is given by the number of data bytes

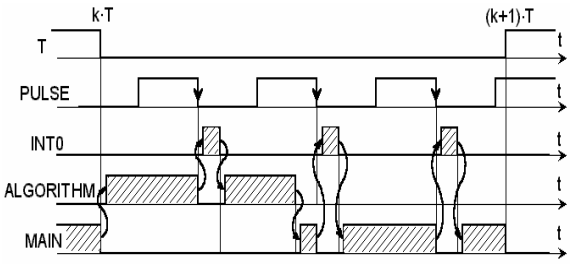



Figure 2: The proposed timing for on-line control based on interupts.

$n_{Bytes}$ saved for each T and the regime duration $\Delta t_{pos}$. A too much little T can lead to an outrunning of the available memory space $V_{data\ mem.}$. Then:

$$T \geq \frac{\Delta t_{pos.\max} \cdot n_{Bytes}}{V_{data\ mem.}} \quad (5)$$

The basic software structure for the on-line control is presented in figure 3. The main program performs an initial preparation of the interrupt system and T2 timer. Then, the interruptible loop makes some auxiliary processing concerning data displaying and a test for the flag F which signals a null position error. Main tasks are those monitored by the hardware interrupt INT0 and the software interrupt ALG / T2. The first is a very fast one and is absolutely the same for all type algorithms, which control the position and speed. A more complex interrupt routine is ALG / T2. The specific tasks concern some additional computing procedures for the control $c_k$ and a characteristic up-to-date for addresses content allocated to regressive variables of standard algorithms. The processing of the obtained control means the extraction of one effective control $c_{k\ out}$ in 8 bit format which acts on a power supply.

## 2.1 Standard Digital (Micro) Controller

The author combined the computations for both loops into a single relation for the control:

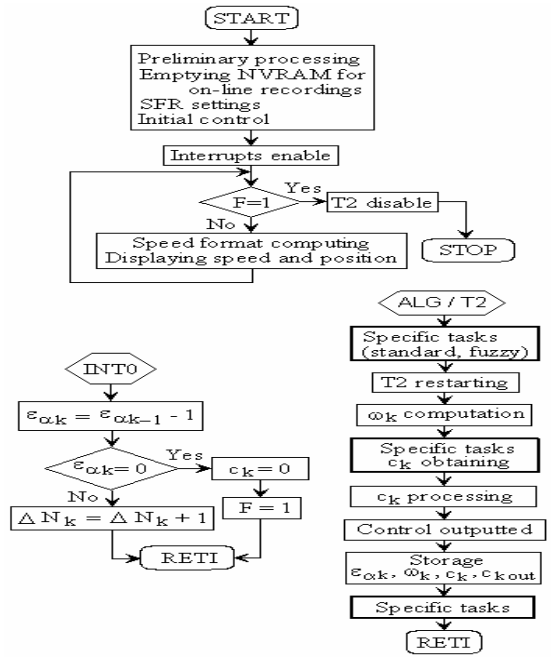



Figure 3: Flow chart of the on-line control program.

$$c_k = c_{k-1} + k_{p\omega} \cdot (\varepsilon_{\omega k} - \varepsilon_{\omega k-1}) + k_{p\omega} \cdot T \cdot \varepsilon_{\omega k} / T_i \quad (6)$$

$k_{p\alpha}$, $k_{p\omega}$ and $T_i$ are the tuning parameters of position and speed loops. Figure 4 has a more detailed systemic structure of the figure 2a. The next form was obtained as an on-line optimal one:

$$c_k = c_{k-1} + A \cdot \Delta N_{k-1} + B \cdot \Delta Nk + C \cdot \varepsilon_{\alpha k} \quad (7)$$

$$A = k_{p\omega} \cdot c_{sp};\ B = -k_{p\omega} \cdot c_{sp} \cdot (k_{p\alpha} + T / T_i)$$
$$C = -k_{p\omega} \cdot k_{p\alpha} \cdot T / T_i \quad (8)$$

A, B and C are the new tuning parameters and all other variables are delivered by INT0. The real-time implementation was prepared by various simulations on a model having the structure and the characteristics very close to the real operation and capabilities of the system and of the microcontroller. The algorithm was applied for simulation and on-line control for a drive system with: a 12 V DC brushed low-inertia motor; an 8 bits microcontroller; an encoder with 2500 lines / rev. A multicriterial optimal conditioning (Mihai, 1999, 2004) of the involved data meant T = 2.456 ms and $c_{sp} = 1.02 \cong 1$ (an ideal value). The figure 5 presents the simulation results by the main macroscopic variables. The on-line results are those from figure 6, the behaviour being like the expected one by simulation. Some differences are in connection with a particular strategy for the pre-final time segment (Mihai, 2004). Figure 7 is a witness of what is happening in real-time, the analyser recordings giving details for the timing (including the T value, interrupt events) and for precise evaluations for each activated task. Notations: SPER-sampling period; PULSE- encoder pulses; INT0-external interrupt service routine; PROC-all speed loop tasks; ARITH- arithmetic routines; SCON-control tasks. Figure 7a reveals the timing for a low speed and the total processing time for having a control: 452 μs. The next diagram is for the rated speed and the real T - 2.47 ms. 7c makes a precise evaluation of the final position error: 2.5 pulses, that meaning 1 / 1000 rev. during 11.37 ms.

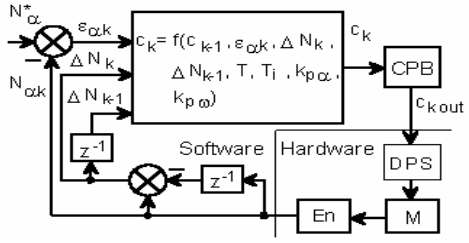



Figure 4: For the standard control algorithm.

Fig. 8 is for T= 2.456 ms (the ideal value) and T = 1ms. It can be seen the worsening of the performance; the control is saturated and a big position overshoot is present. So, a lower T value does not bring a higher performance.

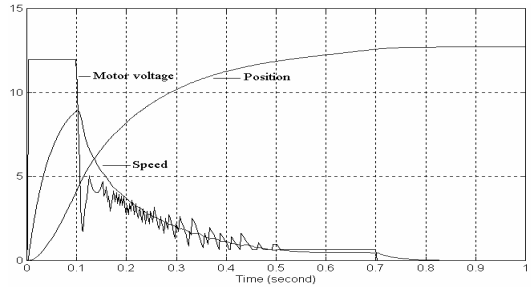



Figure 5: Simulation results / standard algorithm.

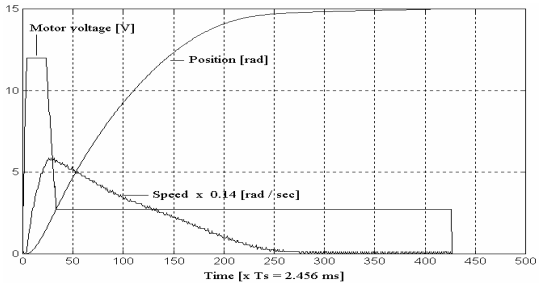



Figure 6: Real-time results / standard algorithm.

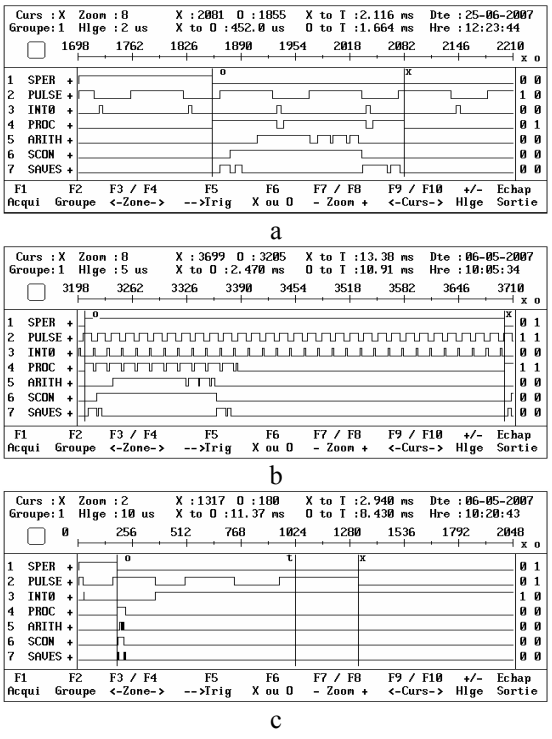



Figure 7: The on-line tasks / standard algorithm.

## 2.2 Standard Algorithm by DSP Controller

The first idea for improving the bad results from fig. 8b is to use a much more performing hardware. The next experiment was made with a brushless DC motor and a DSP controller (Technosoft, 1997). The figure 9a presents the system. The results (in real-time) from figure 9b are for no-load conditions. The figure 9c proves a good general tuning when the motor has a load. This first two result sets were for T=1 ms (speed loop) and T=0.1 ms (current loop). With a faster control sampling (twice), the results are loosing the quality-fig. 9d. A good choice for T is better than an expensive hardware solution.

## 2.3 A Fuzzy Digital Controller

The input variables for the FLC are:

$$\varepsilon_{cank} = (N_\alpha^* - N_{\alpha k}) \cdot \varepsilon_{cank\ max} / N_\alpha^* \geq 0 \qquad (9)$$

$$\Delta\varepsilon_{cank} = \varepsilon_{can\,k} - \varepsilon_{\alpha\,nk-1} = \Delta Nk \cdot \Delta\varepsilon_{cank\,max} \cdot c_{sp} / \Omega_{max} \leq 0 \qquad (10)$$

The most reduced on-line computational effort is obtained using a look-up table (LUT) filled off-line. Fig. 10 gives the image of the equivalent systemic structure of FLC. The control values are stored by the columns concatenation of a matrix with the final control values, so the additional software block CPB is no more necessary. The fuzzy LUT strategy was


Figure 8: T = 2.456 ms (a) and T = 1 ms (b).

applied for the same servodrive as for 2.1. The main characteristic elements are presented by fig. 11. The notations: **az**-almost zero; **vs**-very small; **s**-small; **rs**-relative small; **m**-medium; **b**-big; **vb**-very big. The off-line tuning for the fuzzy rule base was made by simulation on model, with the results from the figure 12. The on-line results are depicted in figure 13 and the on-line timing is given by the figure 14. It can be notice a good concordance between the simulated and the real-time results. The meaning of the additional notations used by the figure 14: NORM-normalization task; LUT-searching in look-up table. The apparent discrepancy between the position smooth evolution and the speed diagram is explained by the fact that the software image $\Delta N_k$ is a truncated value for speed, while the real speed and the position support a filtering effect of mechanical inertia. The ringing of the control is, basically, a result of fuzzy rules commutation. The control variable evolution (motor voltage) reveals a very sensitive behaviour of the controller, better than for the standard controller–the figures 5 and 6. The quality of the whole evolution is also presented in the ($\varepsilon_\alpha$, $\omega$) state-space coordinates, with a good (smooth) behaviour and an entry with very low speed in the proximity of the final point. The same sampling control period as for the standard control algorithm was used: T = 2.456 ms, according to the same multicriterial optimum. It is confirmed the right operation of the 2 interrupts and the existence of quite large time reserve during each T for additional tasks. The diagram from the figure 14a concerns a quasi steady state regime, revealing the typical task distribution. Several values for the tasks time are precisely measured. $\Delta t$ SPER = 2.456 ms << $T_m$ = 50 ms (the electromechanical time constant of the motor); $\Delta t$ PULSE = 125.5 μs - for a speed of 192 RPM. The last diagram caught the final algorithm time segment. The final detectable speed is 3.39 RPM. After the T2 timer (that marks the sampling rate and the cyclic processing for the control determination) is disabled, it can be seen a single pulse coming from the encoder. It is the ideal position error and the real too. Other durations are: $\Delta t$ INT0 = 10 μs; $\Delta t$ PROC = 558 μs < 25 % x T; $\Delta t$ NORM = 310.5 μs; $\Delta t$ LUT = 20.5 μs; $\Delta t$ SAVES = 26 μs (globally). The most of time is necessary for the normalization arithmetic operations.

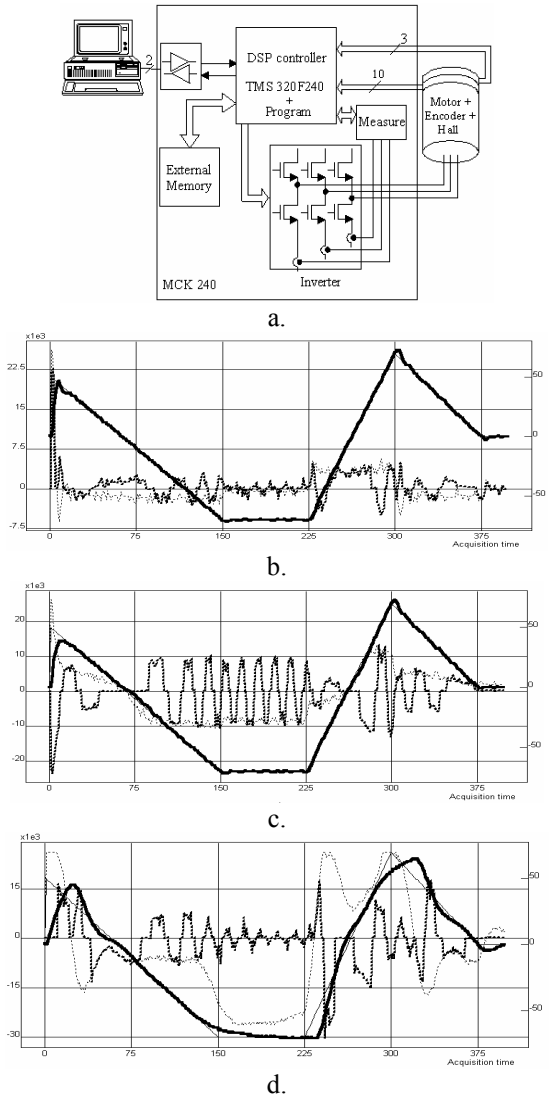



Figure 9: The real-time results for a standard algoritm and a DSP controller.

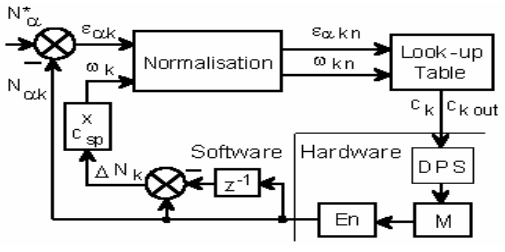



Figure 10: A LUT based FLC.

## 3 CONCLUSIONS

The sampling control period is more than additional tuning parameters for the digital control. The author conceived a multicriterial conditioning relation set with the limitation both for the upper and the lower values for the sampling period. The general timing proposed for the servodrive control is based on imbricate interrupts. A strong real-time constraint is accomplished by a hardware interrupt that makes an acquisition rate different than the control rate. The

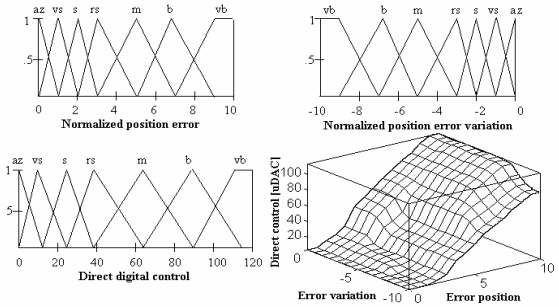



Figure 11: The basic elements for a LUT - FLC.

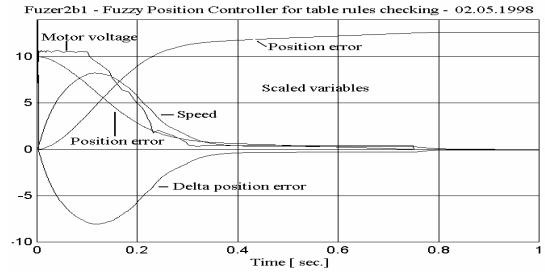



Figure 12: Results for a FLC in simulation.

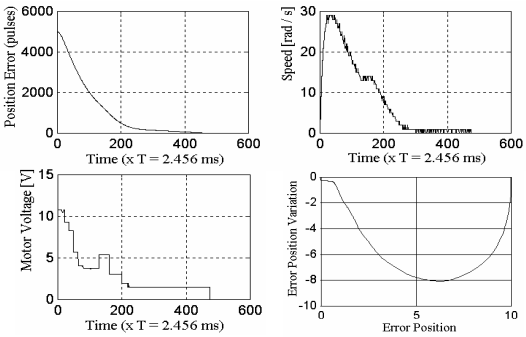



Figure 13: On-line results for a LUT based FLC.

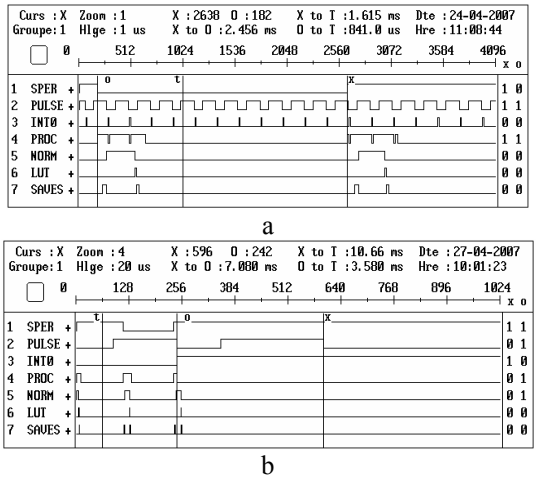



Figure 14: The on-line recordings for a LUT FLC.

on-line results, both for the standard algorithm and the fuzzy control, are very good in terms of the macroscopic variables and for the timing revealed by on-line recordings with a logic analyzer. Some experiments proved that a good choice for the sampling control period is not necessarily related with a high end control processor but is the result for all correlations previously mentioned.

# REFERENCES

Astrom, K.J., Wittenmark, B., 1997. *Computer-controlled systems: theory and design*, Prentice Hall, USA.

Coleman C. P., Godbole D., 1994. A Comparison of Robustness: Fuzzy Logic, PID and Sliding Mode Control*, Proc. of the American Control Conference*, pp. 1654-1659.

Do Wan, K., Jin Bae, P., Young Hoon, J., 2007. Effective digital implementation of fuzzy control systems based on approximate discrete-time models, Automatica *(IFAC Jour.)*, Volume 43 , Issue 10, pp. 1671-1683.

Ibbini, M.S., Jafar, A.S., 2002. Self-Tuning Fuzzy Logic Controller for a Series DC Motor, Proc. *(369) Power and Energy Systems,* Acta Press.

Kozek, M., Lorenz, A., Kampas, Ph., 2007. Modeling and control of an electric servo drive with strong restrictions in the control variable, *Int. Journal of Applied Electromagnetics and Mechanics*, Vol. 25, Number 1-4, pp. 521 – 527.

Mihai, D., Constantinescu C., 1999. Fuzzy Versus Standard Digital Control for a Precise Positioning System with Low-Cost Microcontroller*, PCIM99*, Nurnberg, Proceedings, pp. 249-255.

Mihai, D., 2004. *Systèmes d'entraînements électriques I. Problèmes fondamentaux. Systèmes avec moteurs à courant continu*, Ed. Universitaria, Craiova.

Mihai, D., 2006a. Additional Mathematical Pre-processing for the Fuzzy Control of a Servodrive, *WSEAS Trans. on Circuits and Systems*, Is. 11, Vol. 5, pp. 1575-1580.

Mihai, D., 2006b. An Optimized Fuzzy Control Algorithm for Servodrives. Some Real-Time Experiments. *Proceedings, IS '06,* London, paper 1-4244-0195-X/06-CD ROM, pp. 192-197.

Mrozek, B, Mrozek, Z., 2000. Modelling and Fuzzy Control of DC Drive, *ESM 2000*, Ghent, pp186-190.

Silveira, P. E., Souza, J.R., Biazotto R. de, V. M., 2002. Speed Control of an Autonomous Mobile Robot: Comparison between a PID Control and a Control Using Fuzzy Logic, *J. Braz. Soc. Mech. Sci*., vol.24, no.2, pp.127-129.

Vas, P., 1999. *Artificial-Intelligence-Based Electrical Machines and Drives*, Oxford University Press.

*** Technosoft S. A., 1997, *MCK 240 DSP Motion Control Kit*. User Manual, Switzerland.

# A NEW APPROACH FOR MODELING ENVIRONMENTAL CONDITIONS USING SENSOR NETWORKS

Mehrdad Babazadeh and Walter Lang
*Institute for Microsensors, -actuators and -systems (IMSAS), University of Bremen, Otto-Hahn-Allee, Bremen, Germany*
*mbaba@imsas.uni-bremen.de, wlang@imsas.uni-bremen.de*

Keywords: Temperature, relative humidity, air flow, estimation, grey-box, model.

Abstract: An approach to estimate environmental conditions (ECs), temperature, relative humidity and air flow in a few desired sensor nodes in a wireless sensor network, slept for reducing battery-consumption or inactive due to either empty batteries or out-of-range is presented. A nonlinear, multivariable model containing the interconnections is extracted and using data of surrounding active sensor nodes is broken to the linear models. Unknown parameters of the model are verified by a multivariable identification method. The proposed approach is independent of the type of ventilation system. It can be used in different applications such as designing model base ECs controllers as well as an estimator in fault diagnosis methods.

## 1 INTRODUCTION

Identification, modeling and control of Temperature (*T*), relative Humidity (*H*) and air Flow (*F*) as the environmental conditions (ECs) in the air conditioned closed spaces have gained a lot of attractions during the last few years. Therefore, simple and precise mathematical models can play a key role in these areas. Improving such linear models or proposing new nonlinear models is very vital on this issue. As the first step, we try to achieve a simple mathematical model for the ECs using a wireless sensor network established inside the container loaded with freights. We utilize this model to introduce a new technique to estimate the EC in the place of some desired sensor nodes (DSNs). They may be either in sleep mode or out of service. As stated by the articles, there are three types of models: based on (Sohlberg, 2003), White-box models are made of theoretical considerations where the grey-box models are extracted from the first principles and parameters of the models are obtained by measurement and black-box models are identified only using measurement of the system input and output. The methods achieved to the white, grey and black-box models of *T* for air-handling units have been addressed in (Ghiaus, 2007) , (Shaikh, 2007), (Brecht, 2005), (Desta, 2004), (Frausto 2004). Some other works consider the effects of air flow pattern on the *T* in special cases (Moureh, 2004), (Rouaud, 2002) and (Smale, 2006) is a brief review of numerical models of airflow in refrigerated food applications. (Desta, 2004) outlines a method to achieve an accurate model of *T* in a closed space using both k-ε model and a data-base mechanistic (DBM) modeling technique. It doesn't consider the effect of the heat transfer from the neighboring zones.

All previous models are obtained between input (inlet) and a point of corresponding space. As attested by these methods, the ECs inside the container will change only due to variation in inlet. Some of the models obtained in the existing papers either linear or nonlinear don't consider all of important parameters of the ECs. Furthermore, particular conditions and the limit range of the parameter variations are necessary and despite the high precision, complexity makes them impractical in some applications.

If return to model making in the mentioned space, nonlinear multivariable nature and interconnections between the variables of the ECs in addition to the presence of the freight as an unpredictable, immeasurable disturbance, effects of dynamic of flow, surfaces and walls inside the container increase complexity of the model which we are looking for. Another important factor is disturbance which can be appeared in the different ways and may be cause a big estimation error: (i) Opening the door of the container; (ii) changing either direction or rate of the air flow by some obstacles; (iii) thermal or moisturize influences of

some freight. All attempts in the first step of the present research are towards introducing a grey-box nonlinear model between inlet and one DSN. We will use previous data of a deactivated sensor in addition to the present and previous data of some surrounding sensor nodes to estimate unknown parameters of the related simplified models. According to fig. 1 and also our main proposal in the energy management of the wireless sensor network, there will be a few special key sensor nodes (KSNs) those will send some specific information to main processor and or to the other sensor nodes. The KSNs should be in active mode during the normal operating mode. The KSNs have three major functions: (i) they measure environmental conditions alternatively; (ii) they evaluate measured values and do some estimation of the ECs in the DSNs and update previous models after measuring and receiving some new data; (iii) they will deactivate DSNs when the operational conditions are normal and there are no big changes in the ECs. Usually a while after loading the container, the ECs inside the container have less variations. This duration is the best time to utilize the method to take more DSNs to sleep mode and to estimate the ECS instead of the direct measurement. The KSNs can be located everywhere inside the container, even near the door or near to the inlet. If they are located in some key points, mismatch error due to no considering unpredictable phenomenon will be avoidable because depending on the floating input approach, uncertainties and disturbances are considered indirectly as the input change. It is also independent of the type of the ventilation system. Useful reference for sensor networks is (J. Elson, 2004).

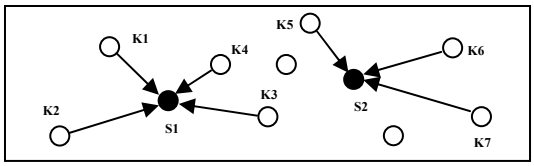



Figure 1: Proposed sensor network.

## 2 PROBLEM FORMULATION

Fig. 2 shows a general scheme of the system, inside the container between the inlet and a spatial position. It is a complicate, time and place dependent, multi-variable system. It consists of three inputs, three outputs, disturbance and noise. Due to the coupling in the ECs, doing independent experiments in the actual container is difficult. It completely depends on the initial conditions so that a change in the *T* or relative humidity of the inlet may change both *T* and *H* in all positions of the space. Variation in the rate of input air flow changes the measurement results and disturbance may change all the results so that based on the existing conditions, measured values might be different even in the same place.

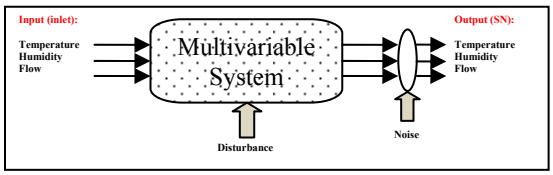



Figure 2: Schematic of Container as a MIMO model.

Floating input approach identifies multivariable models between the KSNs and the DSNs, not between the inlet and a DSN. Every non modeled disturbances which excite some KSNs, is modeled as an implicit input change, not a pure disturbance. Now, the new input nodes (KSNs) in the defined multi-input and single-output (MISO) system change output nodes (DSNs). Fig. 3 shows K1, K2, K3 and K4 as the KSNs and S1 as the DSN.

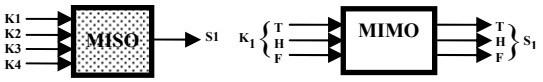



Figure 3: Models between the KSNs and a DSN.

The first step for modeling is using linear transfer function matrix. Without considering noise we have:

$$\begin{pmatrix} T_{SN} \\ H_{SN} \\ F_{SN} \end{pmatrix} = \begin{pmatrix} G_T & -G_{HT} & 0 \\ -G_{TH} & G_H & 0 \\ 0 & 0 & G_F \end{pmatrix} * \begin{pmatrix} T_{in} \\ H_{in} \\ F_{in} \end{pmatrix} \quad (1)$$

($T_{SN}$, $H_{SN}$ and $F_{SN}$) and ($T_{in}$, $H_{in}$ and $F_{in}$) are respectively measured value of (*T*, *H* and *F)* in SN and inlet. Whereas *T* and *H* have opposite effects on each other, we assign negative sign for the interconnection. It is assumed that *F* has no direct effect on the steady state values of *T* and *H,* but it influences on the speed of their variations. However, the effect of *F* are included in all $G_T$, $G_H$, $G_{HT}$ and $G_{TH}$ (which are transfer functions between different parameters of the ECs) with some exponential functions that we will mention later. To investigate validity of the model we employ a reverse lemma and some assumptions in different border conditions.

Assumption 1, steady state values of *T* and *H*:

$$T_{in} = T_{max} \quad , \quad H_{in} = H_{min} = 0 \tag{2}$$

$$T_{Oss} = \lim_{t\to\infty} Z^{-1}(G_T * T_{max}) \tag{3}$$

$$H_{Oss} = -\lim_{t\to\infty} Z^{-1}(G_{TH} * T_{max}) \tag{4}$$

$$\text{Must: } T_{Oss} \le T_{max} \quad , \quad H_{Oss} \ge 0 \xrightarrow{\text{Then}} \tag{5}$$

$$0 \le |G_H|, |G_{TH}| \le 1, |T_{max}| \ge 0 \rightarrow H_{oss} \le 0 \tag{6}$$

It can't be correct because, negative *H* can't be occurred. We consider some permissible margins so that *T* and *H* locate in the mentioned margin:

$$\text{Assumption 2: } T_{in} = T_{max}, H_{in} = H_{min} \neq 0 \tag{7}$$

$$0 \le Z^{-1}(G_H * H_{min} - G_{TH} * T_{max}) \le H_{Omax} \tag{8}$$

$$Z^{-1}\frac{G_H * H_{min} - H_{Omax}}{G_{TH}} \le T_{max} \le Z^{-1}\frac{G_H * H_{min}}{G_{TH}} \tag{9}$$

$$T_{Omin} \le Z^{-1}(G_T * T_{max} - G_{HT} * H_{min}) \le T_{Omax} \tag{10}$$

$$Z^{-1}\frac{G_T * T_{max} - T_{Omax}}{G_{HT}} \le H_{min} \le Z^{-1}\frac{G_T * T_{max} - T_{Omin}}{G_{HT}} \tag{11}$$

$$\text{Assumption 3: } Z^{-1}\frac{G_H * H_{max} - H_{Omax}}{G_{TH}} \le T_{min} \le Z^{-1}\frac{G_H * H_{max}}{G_{TH}} \tag{12}$$

$$T_{Omin} \le Z^{-1}(G_T * T_{min} - G_{HT} * H_{max}) \le T_{Omax} \tag{13}$$

$$Z^{-1}\frac{G_T * T_{min} - T_{Omax}}{G_{HT}} \le H_{max} \le Z^{-1}\frac{G_T * T_{min} - T_{Omin}}{G_{HT}} \tag{14}$$

$$\text{We should have: } T_{max} = Z^{-1}\frac{G_H * H_{min}}{G_{TH}} \tag{15}$$

$$H_{min} = Z^{-1}\frac{G_T * T_{max} - T_{Omax}}{G_{HT}} \tag{16}$$

$$T_{min} = Z^{-1}\frac{G_H * H_{max} - H_{Omax}}{G_{TH}} \tag{17}$$

$$H_{max} = Z^{-1}\frac{G_T * T_{min} - T_{Omin}}{G_{HT}} \tag{18}$$

Having $H_{min}$ and $H_{max}$, other input limitations will be verified. Then there are the specific bands for inputs so that outputs of linear model will be located in the admissible areas. Accordant with the lemma, linear model (1) can't be a proper model. The nonlinear model will be made based on the basic knowledge of the nonlinear nature of the interconnections. Considering some linear transfer functions for direct effects and obtained nonlinear functions for the interactions, we have:

$$\begin{pmatrix} T_{SN} \\ H_{SN} \\ F_{SN} \end{pmatrix} = \begin{pmatrix} G_{T,F} * T_{inlet} + g_{(H,F_{inlet})} + N_T \\ f_{(T,F_{inlet})} + G_{H,F} * H_{inlet} + N_H \\ G_F * F_{inlet} + N_F \end{pmatrix} \tag{19}$$

g(.) and f(.) are nonlinear interconnections between *T* and *H* which are influenced by *F*. As stated by (Ghiaus, 2007) and (Zerihun Desta, 2004), model of *T* can be a first-order transfer function. We also use the effect of the parameters with the same dimensions in the following:

$$G_T = \frac{M_T}{(\alpha_T * S + 1)} \quad , \quad G_H = \frac{M_H}{(\alpha_H * S + 1)} \tag{20}$$

$$\alpha_T = f(Flow) \ , \ \alpha_H = g(Flow) \tag{21}$$

$\alpha_T$ and $\alpha_H$ illustrate speed of the responses and $M_T$ and $M_H$ steady state values of *T* and *H*. They 74have reverse relation with *F*. Then, the further flow rate, the less $\alpha_T$ and $\alpha_H$. The SNs can detecte variations in the ECs showed by ΔT, ΔH and ΔF.

If the position of the SNs is close, we can assume that all models in mentioned MISO system, showed in fig. 3 are independent. It can be considered as several single-input and single-output (SISO) systems. Now, they should be combined using a multivariable identification method. Accordant with the thermodynamic relations, with 10.1 ºC increasing *T*, *H* will be reduced to the half and we have:

$$H = H_0 * 2^{\frac{-(T-T_0)}{10.1}} \quad , \quad T = T_0 - \frac{10.1}{\ln 2} * \ln\frac{H}{H_0} \tag{22}$$

$$T_{SN}(t) = Z^{-1}(G_T * T_{in}) + \Delta T(t) \tag{23}$$

$$\Delta T = g(.) + N_T \ , \quad \Delta H = f(.) + N_H \tag{24}$$

$$\Delta T = \frac{-10.1}{\ln 2} * \ln \frac{Z^{-1}(G_H * H_{in}) + N_H(t)}{Z^{-1}(M_H * H_0)} \quad (25)$$

$$H_{SN}(t) = Z^{-1}(G_H * H_{in}) + \Delta H(t) \quad (26)$$

$$\Delta H = \left[ [2^{\frac{-[Z^{-1}(G_T * T_{in}) + N_T - M_T * T_0]}{10.1}} - 1] * M_H * H_0 \right] + N_H(t) \quad (27)$$

$Z^{-1}$ is unit delay in field of Z transform. It is probable that the amounts of $T_{oss}$ and $H_{oss}$ are changed because of the variation in air flow pattern. However, we consider it on the transfer functions $G_T$ and $G_H$ when running the on-line estimation. From previous results, we will derive a time dependent, nonlinear, multivariable matrix equation and a function of the several KSNs. Uk is a function to obtain the effects of the KSNs on a DSN.

$$\begin{pmatrix} T_{DSN} \\ H_{DSN} \end{pmatrix} = \bigcup_k \begin{pmatrix} T_{KSN(i)} \\ H_{KSN(i)} \end{pmatrix} \quad (28)$$

## 3 SIMULATIONS

Results of the SISO system with initial conditions in the table 1 has been shown in fig.4. It is noted that a part of the parameters such as time constant of *T* in simulations have been inspired of actual behavior of a real experiment and the rest are based on primary assumptions of the authors.

Table 1: Initial conditions for inlet and S1.

| | $T_0$ | $H_0$ | $F_0$ |
|---|---|---|---|
| **inlet** | 10 | 30 | 15 |
| **DSN(S1)** | 9 | 28.5 | 13.5 |

According to fig. 4 Set points of *T* at 2000, *H* at (12000 and 35000) and F at (4000 and 7000) seconds change. An obstacle as a disturbance changes the rate of the air flow and influences on the speed of the responses. However, it will not change the steady state value of the ECs. The initial conditions of *T* and *H* in output are different with those in input (inlet) and after changing *T* in input, output changes slowly to a new equilibrium point because the amount of flow is low in the beginning. At 4000 and 7000 seconds air flow increases respectively to $F_{max}/2$ and $F_{max}$ and immediately the responses of *T* and *H* become faster. When *H* in inlet does not change, *H* in output changes only due to changing T in output. There is a similar story for T in output independent of T in input which varies with the variation of H in output. Dashed curves show the ECs in a desired place inside the container.

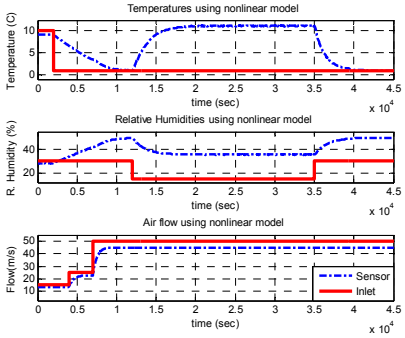



Figure 4: ECs in SN when the ECs in input change.

## 4 AN INDIRECT SOLUTION

We employ the advantages of the sensor network and introduce floating input approach. We assum that m numbers of the KSNs are measuring the conditions when input is inlet and we have:

$$T_{out} = G_T * T_{in} + \Delta T, H_{out} = G_H * H_{in} + \Delta H \quad (29)$$

$$\Delta T = g(.) * G_{dyn(H-T)} + N_T \quad (30)$$

$$\Delta H = f(.) * G_{dyn(T-H)} + N_H \xrightarrow{\text{yields}} \quad (31)$$

$$T_{SN} = T_{linear\ (from\ T)} + T_{non-lin.(from\ H)} \quad (32)$$

$$H_{SN} = H_{linear\ (from\ H)} + H_{non-lin.(from\ T)} \quad (33)$$

We can suppose that the nonlinear interconnections from the inlet are both in the KSNs and the DSNs. Then, we can remove these parts when we consider the KSNs as the input:

$$T_{DSN} = G'_T * T_{KSN}\ ,\ H_{DSN} = G'_H * H_{KSN} \quad (34)$$

$G'_T$, $G'_H$ are the linear transfer functions between a KSN and a DSN and its unknown parameters should be verified using a system

identification technique. Now, we will have some SISO matrix equations which should to be solved. M and P are functions for combining linear effects. We use them in the identification method, indirectly. $U_{Ti}$ and $U_{Hi}$ are new inputs, in the m numbers of the KSNs. $G'_{Ti}$ and $G'_{Hi}$ are linear transfer functions of *T* and *H,* written between the KSNs and the DSN.

$$\begin{pmatrix} T_{DSN} \\ H_{DSN} \end{pmatrix} = \begin{pmatrix} M(G'_{Ti} * U_{Ti}) & 0 \\ 0 & P(G'_{Hi} * U_{Hi}) \end{pmatrix} \quad (35)$$

# 5 RESULTS

As an example, showed in fig. 5, there are two KSNs and one DSN attached to the walls, there are some obstacles so that the change-rate of the ECs near to the SNs is different with those in inlet. There are also different amounts of initial conditions for different SNs because of their positions or corresponding measurement errors (table 2). The simulation results has been shown in fig. 6.

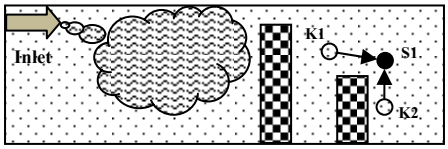



Figure 5: A container with inlet, KSNs and DSN.

Table 2: Initial conditions.

| | $T_0$ | $H_0$ | $F_0$ |
|---|---|---|---|
| inlet | 10 | 30 | 15 |
| K1 | 9 | 28.5 | 13.5 |
| K2 | 8.5 | 27 | 3 |
| S1 | 8 | 25.5 | 10 |

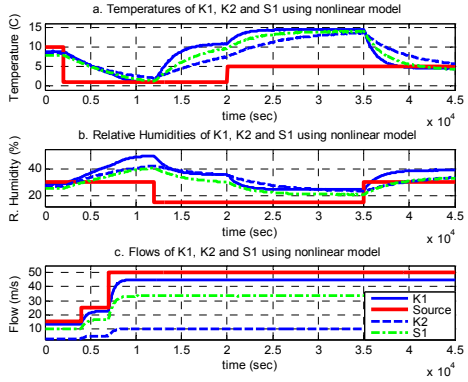



Figure 6: Outputs when *T*, *H* and *F* in input change.

As shown in fig. 6, curves of K1, K2 and S1 are according with the data extracted from models introduced in equations (23) and (26) and curves related to the inlet are the set points. The relations of *T*, *H* and interconnections are updated based on the amount of *F* at the related instant of the simulation.

# 6 OFF-LINE IDENTIFICATION

Refer to equation (35), there are separate MISO systems for *T* as well as *H*. All unknown parameters should be determined using an off-line identification technique. Then, we assume that KSNs are active and there is a failure on the DSN or it is in sleep mode and having new inputs we will have the new estimations of the ECs in the DSNs using existing transfer functions. The temperature estimation results have been shown in fig. 7 and fig. 8 with the SISO and MISO models, respectively. To show capability of the method, the results have been plotted together with the previous results of the EC in S1 from introduced nonlinear model. the measured *T* of K1 in the vicinity of S1, without any variation in *T* of inlet and K2. We obtain its effects on S1 when estimated by K1 and K2 compare with a regular estimation method using model obtained from inlet-DSN. The Solid wide curves illustrates nonlinear model output and dashed curves represent obtained results separately using linear models and then with MISO estimation using output error (OE) method in system identification toolbox of Matlab:

$$\frac{B_i(Z)}{F_i(Z)} = \frac{(b_0 + b_1 * Z^{-1} \dots + b_m * Z^{-m})}{(1 + a_1 * Z^{-1} \dots + a_n * Z^{-n})} \quad (36)$$

$$y(t) = \sum_{i=1}^{nu} \frac{B_i(Z)}{F_i(Z)} u_i(t - nk_i) + e(t) \quad (37)$$

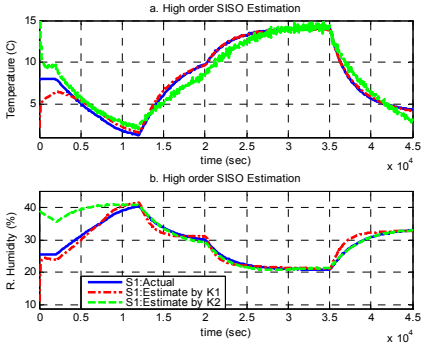



Figure 7: Actual and estimated *T* and *H*, model with the order three using K1 and K2, separately.

To achieve a desired speed and regard to the nonlinear nature of the responses that we have still in the SNs, we utilize a linear transfer function with the order more than two. Whereas the higher order models will cause some difficulties in the application, we don't use the order more than three. Separate estimations using SISO models have less accuracy than those using MISO models.

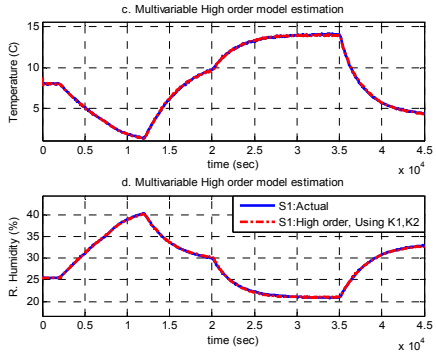



Figure 8: *T* and *H*, actual, estimation using high order multivariable model from K1, K2.

More important results are obtained when there is a disturbance in vicinity of the SNs influences some of the KSNs. fig. 9 shows the variation of *T* at 25000 seconds which affects both K1 and S1. Model obtained from inlet-S1 can't show this influence on S1 because there are no influences on the inlet. However, floating input method can estimate it because at least one KSN senses it.

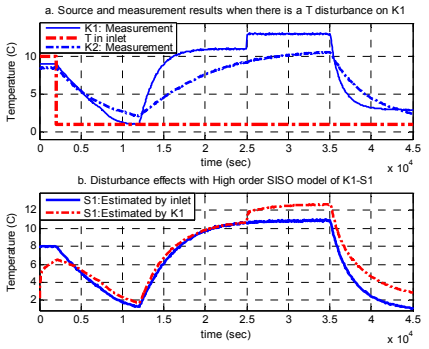



Figure 9: a. measured *T* in inlet, K1 and K2 and b. estimation using inlet and K1with existing a disturbanc.

# 7 CONCLUDING REMARKS

This paper propuses a new hybrid model for environmental conditions inside a container and shows that it has much more accuracy for wide range of parameter variations compared to other conventional linear models between inlet and a desired place. The new technique provides a simplified multivariable model based on the surrounding sensor nodes used for estimating the ECs in the desired nodes. The simulation results and mathematical proofs for different situations endorse the capability of the proposed technique. At the end, it should be noted that the comparison among different multivariable estimation methods and their implementations as well as finding the minimum number and the best place of the KSNs are real challenges main concerns on this issue.

# REFERENCES

Ghiaus, C., Chicinas, A. and Inard, C., April 2007, "Grey-box identification of air-handling unit elements, *Control Engineering Practice*", Vol.15, Issue 4, pp. 421-433.

Shaikh, N. I and Prabhu, V., May 2007,"Mathematical modeling and simulation of cryogenic tunnel freezers", *Journal of Food Engineering*, Vol. 80, Issue 2, pp 701-710.

Smale, N.J., Moureh, J. and Cortella, G., Sep. 2006, "A review of numerical models of airflow in refrigerated food applications", *International Journal of Refrigeration*, Vol. 29, Issue 6, pp. 911-930.

Van Brecht, A., Quanten, S., Zerihundesta, T., Van Buggenhout, S. and Berckmans, D., 20 Jan. 2005, "Control of the 3-D spatio-temporal distribution of air temperature", *International Journal of Control*, 78:2, pp. 88- 99.

Zerihun Desta, T., Van Brecht, A., Meyers, J., Baelmans, M. and Berckmans, D., June 2004, "Combining CFD and data-based mechanistic (DBM) modeling approaches, Energy and Buildings", Vol. 36, Issue 6, pp 535-542.

Frausto, H. U. and Jan G. Pieters, Jan.2004, "Modeling greenhouse temperature using system identification by means of neural networks", *Neurocomputing*, Vol. 56, pp.423-428.

J. Moureh and Flick, D., Aug. 2004; "Airflow pattern and temperature distribution in a typical refrigerated truck configuration loaded with pallets", *International Journal of Refrigeration*, V. 27, Issue 5, pp. 464-474.

Rouaud, O. and Havet, M., May 2002, "Computation of the airflow in a pilot scale clean room using K-ε turbulence models", *International Journal of Refrigeration*, Vol. 25, Issue 3, pp. 351-361.

Sohlberg B., May 2002, Apr. 2003, "Grey box modeling for model predictive control of a heating process", *Journal of Process Control*, Vol. 13, Issue 3, pp. 225-238.

J. Elson and D. Estrin. Sensor Networks: A Bridge to the Physical World, chapter 1. Wireless Sensor Networks. Kluwer Academic Publishers, 2004.

# PASSIVITY OF A CLASS OF HOPFIELD NETWORKS
## *Application to Chaos Control*

Adrian–Mihail Stoica
*Faculty of Aerospace Engineering, University "Politehnica" of Bucharest, Str. Polizu, No. 1, Ro-011061, Bucharest, Romania*
*amstoica@rdslink.ro*

Isaac Yaesh
*Control Department, IMI Advanced Systems Div., P.O.B. 1044/77, Ramat–Hasharon, 47100, Israel*
*iyaesh@imi-israel.com*

Keywords: Chaos, Stochastic systems, Hopfield Neural Networks, Recurrent Neural Networks, Sprocedure, Linear matrix inequalities, Direct Adaptive Control.

Abstract: The paper presents passivity conditions for a class of stochastic Hopfield neural networks with state–dependent noise and with Markovian jumps. The contributions are mainly based on the stability analysis of the considered class of stochastic neural networks using infinitesimal generators of appropriate stochastic Lyapunov–type functions. The derived passivity conditions are expressed in terms of the solutions of some specific systems of linear matrix inequalities. The theoretical results are illustrated by a simplified adaptive control problem for a dynamic system with chaotic behavior.

## 1 INTRODUCTION

Hopfield networks are symmetric recurrent neural networks which exhibit motions in the state space which converge to minima of energy.

Symmetric Hopfield networks can be used to solve practical complex problems such as implement associative memory, linear programming solvers and optimal guidance problems. Recurrent networks which are non symmetric versions of Hopfield networks play an important role in understanding human motor tasks involving visual feedback (see (Cabrera and Milton, 2004) - (Cabrera et al., 2001) and the references therein). Such networks seem to be subject to effects of state-multiplicative noise, pure time delay (see (Hu et al., 2003), (X. Liao and Sanchez, 2002) and (Stoica and Yaesh, 2006)) and even multiple attractors which can be caused by Markov jumps. Even without Markov jumps, a non symmetric class of Hopfield networks is able to generate chaos (Kwok et al., 2003). Therefore, Hopfield networks can be used (Poznyak and Sanchez, 1999) as identifiers of unknown chaotic dynamic systems. The resulting identifier neural networks have been used in (Poznyak and Sanchez, 1999) to derive a locally optimal robust controller to remove the chaos in the system.

In this paper, we consider to replace the robust controller of (Poznyak and Sanchez, 1999) by a direct adaptive controller. More specifically we consider the so called Simplified Adaptive Control (SAC) method (Kaufman et al., 1998) which applies a simple proportional controller whose gain is adapted according the squared tracking error. Since such controllers' stability proof involves a passivity condition, we derive a passivity result for the Hopfield network. Our results are, in fact, developed for a generalized version of non symmetric Hopfield networks including Markov jumps of the parameters and state multiplicative noise thus allowing a wider stochastic class of chaos generating systems to be considered.

The paper is organized as follows. In Section 2, the problem is formulated and in Section 3 Linear Matrix Inequalities (LMIs) based conditions are derived for passivity analysis. In Section 4 a chaos control example is given and finally Section 5 includes concluding remarks.

Throughout the paper $\mathcal{R}^n$ denotes the $n$ dimensional Euclidean space, $\mathcal{R}^{n\times m}$ is the set of all $n \times m$ real matrices, and the notation $P > 0$, (respectively, $P \geq 0$) for $P \in \mathcal{R}^{n\times n}$ means that $P$ is symmetric and positive definite (respectively, semi-definite). Throughout the paper $(\Omega, \mathcal{F}, \mathcal{P})$ is a given probability space; the argument $\theta \in \Omega$ will be suppressed. Expectation is denoted by $E\{.\}$ and conditional expectation of $x$ on the event $\theta(t) = i$ is denoted by $E[x|\theta(t) = i]$.

## 2 PROBLEM FORMULATION

The neural network proposed by Hopfield, can be described by an ordinary differential equation of the form

$$\dot{v}_i(t) = a_i v_i(t) + \sum_{j=1}^{n} b_{ij} g_j(v_j(t)) + \bar{c_i} = \kappa_i(v), 1 \leq i \leq n \tag{1}$$

where $v_i$ represents the voltage on the input of the $i$th neuron, $a_i < 0, 1 \leq i \leq n$, $b_{ij} = b_{ji}$ and the activations $g_i(\cdot), i = 1,...,n$ are $C^1$–bounded and strictly increasing functions.

This network is usually analyzed by defining the network energy functional:

$$\begin{aligned} E(v) &= -\sum_{i=1}^{n} a_i \int_0^{v_i} u \frac{dg_i(u)}{du} du \\ &\quad -\frac{1}{2}\sum_{i,j=1}^{n} b_{ij} g_i(v_i) g_j(v_j) - \sum_{i=1}^{n} \bar{c_i} g_i(v_i) \end{aligned} \tag{2}$$

where it can be seen that $\frac{dE}{dt} = -\sum \frac{dg_i(v_i)}{dv_i} \kappa_i(v)^2 \leq 0$ where the zero rate of the energy is obtained only in the equilibrium points, also referred to as attractors, where

$$\kappa_i(v^0) = 0, 1 \leq i \leq n \tag{3}$$

The network is then described in matrix form as:

$$\dot{v}(t) = Av(t) + Bg(v) + \bar{C}, 1 \leq i \leq n \tag{4}$$

where

$$A := diag(a_1,...,a_n), B := [b_{ij}]_{i,j=1,...,n},$$
$$\bar{C} := \begin{bmatrix} \bar{c_1} & \bar{c_2} & ... & \bar{c_n} \end{bmatrix}^T, v := \begin{bmatrix} v_1 & v_2 & ... & v_n \end{bmatrix}^T$$

and where

$$g(v) := \begin{bmatrix} g_1(v_1) & g_2(v_2) & ... & g_n(v_n) \end{bmatrix}^T$$

The stochastic version of this network driven by white noise, has been considered in (Hu et al., 2003) where the stochastic stability of (1) has been analyzed and where it has been shown that the network is almost surely stable when the condition $\frac{dE}{dt} \leq 0$ is replaced by $\angle E \leq 0$ where $\angle$ is the infinitesimal generator associated with the Itô type stochastic differential equation (4). This condition has been shown in (Hu et al., 2003) to be satisfied only in cases where the driving noise in (1) is not persistent. This non persistent white noise can be interpreted as a white-noise type uncertainty in $A$ and $B$, namely state-multiplicative noise. In (Stoica and Yaesh, 2005)-(Stoica and Yaesh, 2006) Hopfield networks with Markov jump parameters have been considered to represent also non zero mean uncertainties in these matrices. Encouraged by the insight gained in (Cabrera and Milton, 2004) and (Cabrera et al., 2001) regarding the role of state-multiplicative noise and time delay (see also (Mazenc and Niculescu, 2001)) in visuo-motor control loops, we generalize the results of (Stoica and Yaesh, 2005) to include this effect. The Lur'e - Postnikov systems approach ((Lure and Postnikov, 1944), (Boyd et al., 1994)) is invoked to analyze stability and disturbance attenuation (in the $H_\infty$ norm sense) and the results are given in terms of Linear Matrix Inequalities (LMI).

To analyze input output properties we first define the error of the Hopfield network output with respect to its equilibrium points by

$$x(t) = v(t) - v^0. \tag{5}$$

and assume that the errors vector $x(t)$ satisfy

$$\begin{aligned} dx = (A_0(\theta(t))x + B_0(\theta(t))f(y) + D(\theta(t))u(t))dt \\ +A_1(\theta(t))xd\eta + B_1(\theta(t))f(y)d\xi \end{aligned} \tag{6}$$

where the system measured output is

$$z = L(\theta(t))x + N(\theta(t))u \tag{7}$$

and where

$$y = C(\theta(t))x \tag{8}$$

Note that (6) was obtained from (4) by replacing $Adt$ by $A_0dt + A_1d\xi$, $Bdt$ by $B_0dt + B_1gd\xi$ and $f(x) = g(x + v_0) - g(v_0)$. The control input $u(t)$ as introduced to provide a stochastic version of (Poznyak and Sanchez, 1999) allowing the considered Hopfield network to serve also a chaotic system identifier. We note that (Poznyak and Sanchez, 1999) the control signal is $u = \phi(r)u$ rather than just $u$ where $\phi(r)$ is a diagonal matrix having $f_i(r_i)$ on its diagonal, where $r = H(\theta(t))x$. We have taken for simplicity $\phi = I$ which is also motivated by our example in Section IV.

Note also that the matrices $A_0(\theta(t)), A_1(\theta(t)), B_0(\theta(t)), B_1(\theta(t)), D(\theta(t)), C_1(\theta(t)), C_2(\theta(t))$ and $L(\theta(t))$ are piecewise constant matrices of appropriate dimensions whose entries are dependent upon the mode $\theta(t) \in \{1,...,r\}$ where $r$ is a positive integer denoting the number of possible models between which the Hopfield network parameters can jump. Namely, $A_0(\theta(t))$ attains the values of $A_{0,1}, A_{0,2}, ..., A_{0,r}$, etc. It is assumed that $\theta(t), t \geq 0$ is a right continuous homogeneous Markov chain on $\mathcal{D} = \{1,...,r\}$ with a probability transition matrix

$$\begin{aligned} &P(t) = e^{Qt}; Q = [q_{ij}]; q_{ii} < 0; \textstyle\sum_{j=1}^{r} q_{ij} = 0; \\ &i = 1,2,..,r. \end{aligned} \tag{9}$$

Given the initial condition $\theta(0) = i$, at each time instant $t$, the mode may maintain its current state or jump to another mode $i \neq j$. The transitions between

the $r$ possible states, $i \in \mathcal{D}$, may be the result of random fluctuations of the actual network components (*i.e.* resistors, capacitors) characteristics or can used to artificially model deliberate jumps which are the result of parameter changes in an optimization problem the network is used to solve. In visuo-motor tasks one may conjecture that proportional and derivative feedbacks are applied on the basis of time sharing, where transition probabilities define the statistics of switching between the two modes. Although there is no evidence for this conjecture, the model analyzed in the present paper can be used to check its stability and $L_2$ gain.

In the forthcoming analysis, we will assume that the components $f_i, i = 1,...,n$ of $f(\xi)$ are assumed to satisfy the sector conditions

$$0 \leq \zeta_i f_i(\zeta_i) \leq \zeta_i^2 \sigma_i \tag{10}$$

which are equivalent to

$$-F_i(\zeta_i, f_i) := f_i(\zeta_i)(f_i(\zeta_i) - \sigma_i \zeta_i) \leq 0 \tag{11}$$

We shall further assume that

$$\frac{\partial f_i}{\partial \zeta_i} \leq \sigma_i, \quad i = 1,...,n. \tag{12}$$

Although the latter assumption of (12) further restricts the sector–type one class of (11), the applicability of our results remains since it is fulfilled by the usual nonlinearities as saturation, sigmoid, etc., used in the neural networks.

Some additional notations are now in place. We define

$$S = diag\left( \ \sigma_1, \sigma_2, \quad ... \quad , \sigma_n \ \right)$$

where $\sigma_i$ are the nonlinearity gains of (11).

As mentioned above we shall analyze passivity (in stochastic sense) conditions for the systems (6)-(8b) which is expressed as:

$$J = E\left\{ \int_0^\infty \left( z^T(t)\, u(t) \right) dt \right\} > 0, \; x(0) = 0. \tag{13}$$

## 3 PASSIVITY ANALYSIS

Introduce the Lyapunov–type function:

$$\begin{array}{c} V(x(t), \theta(t)) = x^T(t) P(\theta(t)) x(t) \\ +2\sum_{k=1}^n \lambda_k \int_0^{C_k x} f_k(s)\, ds. \end{array} \tag{14}$$

depending on the nonlinearities $f_i(y_i) = f_i(C_i x)$ via the constants $\lambda_i$ where $C_i$ is the $i$'th row in $C$. As it was mentioned in (Boyd et al., 1994), $V$ of (14) defines a parameter-dependent Lypaunov function. To see this, consider the simple case of $f_i(x_i) = x_i \sigma_i$ and get $V(x, \sigma_1, \sigma_2, ..., \sigma_n) = x^T (P + S^{\frac{1}{2}} \Lambda S^{\frac{1}{2}}) x$ which depends on the parameters $\sigma_i, i = 1,2,..,n$ and on the constants $\lambda_i, i = 1,2,..,n$ via (18) in the sequel. Applying the Itô–type formula (see (Dragan and Morozan, 1999), (Dragan and Morozan, 2004) and (Fen et al., 1992)) for $V(x, \theta)$ it follows that:

$$\begin{array}{c} E\left\{ V(x, \theta(t) \,|\theta(0)) \right\} - E\left\{ V(0, \theta(0) \,|\theta(0)) \right\} = \\ E\left\{ \int_0^t \mathcal{L} V(x, \theta(s))\, ds \right\} \end{array}$$

where

$$\begin{array}{l} \mathcal{L} V(x, \theta) := (A_0(\theta) x + B_0(\theta) f(\underline{y}) + D(\theta) u)^T \frac{\partial V}{\partial x} \\ + x^T A_1^T(\theta) \bar{P} A_1(\theta) x + f^T B_1^T(\theta) \bar{P} B_1(\theta) f \\ + \sum_{j=1}^r q_{ij} x^T P_j x. \end{array} \tag{15}$$

where

$$\bar{P}(\theta, \lambda_1, \lambda_2, ..., \lambda_n) := P(\theta) + diag\left( \lambda_1 \frac{\partial f_1}{\partial x_1}, ..., \lambda_n \frac{\partial f_n}{\partial x_n} \right),$$

with the dependence on its arguments being omitted and where for simplicity we have used the notation $f := f(y(t))$. Then the condition (13) is fulfilled if

$$\mathcal{L} V < 2 z^T u \tag{16}$$

which becomes:

$$\begin{array}{l} \left( x^T A_{0i}^T + f^T B_{0i}^T + u^T D_i^T \right) \left( P_i x + C^T \Lambda f \right) \\ + \left( x^T P_i + f^T \Lambda C \right) \left( A_{0i} x + B_{0i} f + D_i u \right) \\ + x^T A_{1i}^T \bar{P}_i A_{1i} x + f^T B_{1i}^T \bar{P}_i B_{1i} f \\ + \sum_{j=1}^r q_{ij} x^T P_j x - u^T L_i x - x^T L_i^T u \\ - u^T (N_i + N_i^T) u < 0, \quad i = 1,...,r, \end{array} \tag{17}$$

where $\bar{P}_i$ denotes $\bar{P}(\theta = i, \lambda_1, \lambda_2, ..., \lambda_n)$ and where

$$\Lambda := diag\left( \ \lambda_1, \lambda_2, \quad ... \quad , \lambda_n \ \right). \tag{18}$$

In order to explicitly express (17), the assumption (12) will be used. Indeed, using the inequalities (12) it follows that conditions (17) are satisfied if the following inequalities are satisfied:

$$\begin{array}{l} -F_{i0}(x, f) := \\ x^T \left[ A_{0i}^T P_i + P_i A_{0i} + A_{1i}^T \left( P_i + C^T S^{\frac{1}{2}} \Lambda S^{\frac{1}{2}} C \right) A_{1i} \right. \\ \left. + L_i^T L_i + \sum_{j=1}^r q_{ij} P_j \right] x + f^T \left[ B_{0i}^T C^T \Lambda + \Lambda C B_{0i} \right. \\ \left. + B_{1i}^T \left( P_i + C^T S^{\frac{1}{2}} \Lambda S^{\frac{1}{2}} C \right) B_{1i} \right] f \\ + f^T \left( B_{0i}^T P_i + \Lambda C A_{0i} \right) x + x^T \left( P_i B_{0i} + A_{0i}^T C^T \Lambda \right) f \\ - u^T (L_i - D_i^T P_i) x - x^T (L_i^T - P_i D_i) u \\ - u^T (N_i + N_i^T) u < 0. \end{array} \tag{19}$$

Using the $\mathcal{S}$-procedure (*e.g.* (Boyd et al., 1994)) one, therefore, obtains that (16) subject to (11) is satisfied if there exist $\tau_i \geq 0, i = 1,2,...,n$ so that

$$F_{i0}(x, f) - \sum_{k=1}^n \tau_k F_k(x, f) \geq 0. \tag{20}$$

Denoting

$$T := diag\left( \tau_1, \tau_2, \quad ... \quad , \tau_n \right) \tag{21}$$

and noticing that

$$\begin{aligned} -\sum_{k=1}^{n} \tau_k F_k(x,f) &= \sum_{k=1}^{n} \tau_k f_k^2 - \tau_k \sigma_k f_k y_k \\ &= f^T T f - \frac{1}{2} f^T TCSx \\ &\quad -\frac{1}{2} x^T SC^T T f, \end{aligned}$$

we get from (20) that:

$$\begin{array}{l} x^T Z_{i11} x + f^T Z_{i12}^T x + x^T Z_{i12} f + f^T Z_{i22} f \\ -u^T (L_i - D_i^T P_i) x - x^T (L_i^T - P_i D_i) u \\ -u^T (N_i + N_i^T) u < 0, \;\; i = 1, ..., r. \end{array}$$

where

$$\begin{aligned} Z_{i11} := &\quad A_{0i}^T P_i + P_i A_{0i} + A_{1i}^T \hat{P}_i A_{1i} + \sum_{j=1}^{r} q_{ij} P_j \\ Z_{i12} := &\quad P_i B_{0i} + A_{0i}^T C^T \Lambda + \frac{1}{2} SC^T T \\ Z_{i22} := &\quad B_{0i}^T C^T \Lambda + \Lambda C B_{0i} + B_{1i}^T \hat{P}_i B_{1i} - T \end{aligned} \tag{22}$$

where

$$\hat{P}_i = P_i + C^T S^{\frac{1}{2}} \Lambda S^{\frac{1}{2}} C \tag{23}$$

These conditions are fulfilled if:

$$\begin{bmatrix} Z_{i11} & Z_{i12} & P_i D_i - L_i^T \\ Z_{i12}^T & Z_{i22} & 0 \\ D_i^T P - L_i & 0 & -(N_i^T + N_i) \end{bmatrix} < 0, \tag{24}$$

$i = 1, ..., r$, with the unknown variables $P_i, \Lambda$ and $T$.

The above developments are concluded in the following result:

***Theorem 1.*** *The system (6)–(7) is stochastically stable and strictly passive if there exist the symmetric matrices $P_i > 0, i = 1, ..., r$, and the diagonal matrices $\Lambda > 0$ and $T > 0$ satisfying the system of LMIs (24) with the notations (22)-(23).* □

# 4 SIMPLIFIED ADAPTIVE CONTROL

In this section we show that the system of (6)-(8) should be regulated using a direct adaptive controller of the type:

$$u = -Kz \tag{25}$$

where

$$\dot{K} = zz^T. \tag{26}$$

Since this type of adaptive control is well–known in the deterministic case (see *e.g.* (Kaufman et al., 1998)), we shall just give a sketch of the proof, emphasizing the particularities arising in the stochastic framework (see also (Yaesh and Shaked, 2005)) analyzed in this paper. We first note that the system (6)-(8) is strictly passive when the passivity condition of (24) is satisfied with $N_i = \varepsilon I$ for $\varepsilon$ that tends to zero. The latter is satisfied (see e.g. (Yaesh and Shaked, 2005) and (Kaufman et al., 1998)) if there exist the symmetric matrices $P_i > 0, i = 1, 2, ..., r$ such that

$$Z_i < 0 \text{ and } i = 1, 2, ..., r, \tag{27}$$

and

$$P_i D_i = L_i^T, i = 1, 2, ..., r, \tag{28}$$

where $Z_i = \begin{bmatrix} Z_{i11} & Z_{i12} \\ Z_{i12}^T & Z_{i22} \end{bmatrix}$. The stochastic closed-loop system obtained from (6), (8) with $u = Kz$ can be written as:

$$\begin{aligned} dx &= [(A_0(\theta(t)) - D(\theta(t)) K_e L(\theta) x) \\ &\quad + B_0(\theta(t)) f(y) + D(\theta(t)) \bar{u}] dt \\ &\quad + A_1(\theta(t)) x d\eta + B_1(\theta(t)) f(y) d\xi \\ z &= L(\theta(t)) x. \end{aligned} \tag{29}$$

where $\bar{u} = -(K - K_e) z$. The above equations hold for any $K_e$ of appropriate dimensions but in the following it will be assumed that $K_e$ is a constant gain for which the system (29) is stochastically passive (some authors call in this case the open-loop system *almost passive*-AP). Note that $K_e$'s existence is needed just for stability analysis and but its value is not utilized in the implementation. In our stochastic context the stochastic stability of this direct adaptive controller (25), (26) (which usually referred to as *simplified adaptive control*-SAC) will be guaranteed by the stochastic version of the AP property. To this end, as in (Kaufman et al., 1998) we will choose the following generalization of the Lyapunov function of (14) to prove the closed-loop stability:

$$\begin{array}{rl} \mathcal{V}(x(t), K(t), \theta(t)) = V(x(t), \theta(t)) \\ + \quad tr(K(t) - K_e)^T (K(t) - K_e) \end{array} \tag{30}$$

where $tr$ denotes the trace and $V$ has the expression (14) with $P(i)$, $i = 1, ..., r$ satisfying the conditions of form (27) and (28) written for the passive system (29) relating $\bar{u}$ and $z$. Then, direct computations show that the infinitesimal generator of $\mathcal{V}$ of the form (30) along the trajectory (29) and subject to the conditions (28) has the expression:

$$\mathcal{L}\mathcal{V}(x(t), K(t), \theta(t) = i) = \bar{x}^T Z_i \bar{x} + 2tr\left( \bar{K}^T \dot{\bar{K}} - \bar{K}^T z z^T \right) \tag{31}$$

where $\bar{K} := K - K_e$ and $\bar{x} = \begin{bmatrix} x \\ f \end{bmatrix}$. Since the system (29) was assumed passive (i.e. (27) is satisfied with

$A_{0i} - D_i K_e L_i$ replacing $A_{0i}$) it follows that $\mathcal{L}V < 0$ and then, choosing $\dot{K} = zz^T$ it results that $\mathcal{L}\mathcal{V} < 0$ which proves the stochastic stability of the resulting closed-loop system. □

We next apply this result in a chaos control problem.

## 5 EXAMPLE - CHAOS CONTROL

Consider a slightly modified version of the third order chaos generator model of (Kwok et al., 2003) described by (6)-(8), where

$$A_0 = \begin{bmatrix} -\varepsilon & 1 & 0 \\ 0 & -\varepsilon & 1 \\ a_1 & a_2 & a_3 \end{bmatrix}, B_0 = D = L^T = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix},$$
$$A_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \sigma \end{bmatrix}, B_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$
$$C^T = \begin{bmatrix} \beta \\ 0 \\ 0 \end{bmatrix}, \tag{32}$$

where $a_1 = -2, a_2 = -1.48, a_3 = -1, \sigma = 0.1, \varepsilon = 0.01$ and $\beta = 10$. The nonlinearity is $f(y) = \alpha tanh(y)$ where $\alpha = 1$.

To establish stability we verify (27)-(28) with $K_e = 10^5$ and find using YALMIP (Löfberg, 2004a)-(Löfberg, 2004b) where $A_0$ is replaced by $A_0 - DK_eL$. Therefore, by the results of Section 4 above, the closed-loop system with the controller (25), (26) is expected to be stochastically stable.

Next we simulate the above system for $500sec$ with an integration step of $0.001sec$ with $u = 0$ for $t \leq 250sec$ and with the SAC controller $u = -Kz$ where $\dot{K} = z^2$ in the rest of the time. The results are given in Fig. 1 - 3 : the phase-plane (i.e. $x_1$ versus $x_2$) trajectories are depicted in Fig. 1, the components $x_i, i = 1,2,3$ of the state-vector and the control input are depicted in Fig. 2, and the adaptive gain $K$ is depicted in Fig. 3. It is seen from these figures that the chaotic behavior characterizing the system in open-loop, is replaced by a stable trajectory at $t \geq 250sec$ where the SAC is applied.

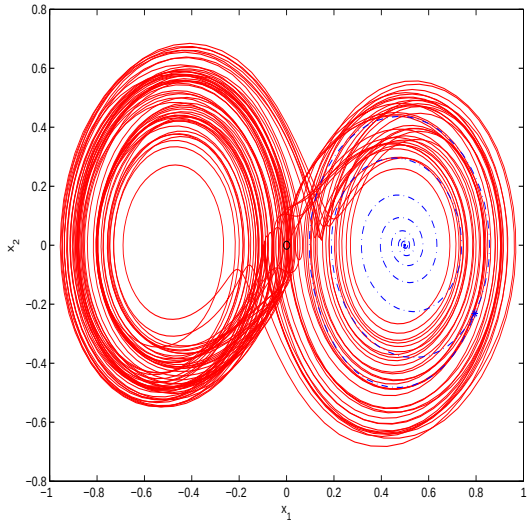



Figure 1: Simulation Results.

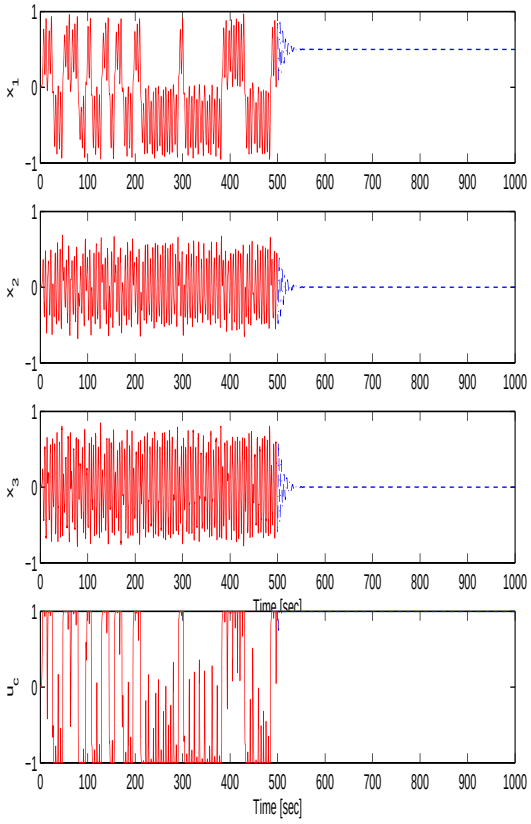



Figure 2: Simulation Results.

## 6 CONCLUSIONS

A class of stochastic Hopfield networks subject to state-multiplicative noise where the network weights jump according a Markov chain process have been considered. Stochastic passivity conditions for such systems have been derived in terms of Linear Matrix Inequalities. The results have been illustrated via simplified adaptive control of a dynamic system which exhibits a chaotic behavior when its is not controlled.

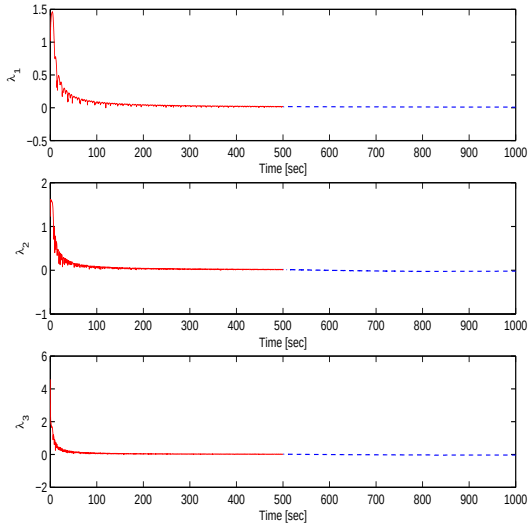



Figure 3: Simulation Results.

The control efficiency in stabilizing the chaotic process has been demonstrated with simulations. The results of this paper should encourage further study of attempts to control chaotic systems with simplified adaptive controllers.

## REFERENCES

Boyd, S., El-Ghaoui, L., Feron, L., and Balakrishnan, V. (1994). *Linear matrix inequalities in system and control theory*. SIAM.

Cabrera, J., Bormann, R., Eurich, C., Ohira, T., and Milton, J. (2001). State-dependent noise and human balance control. *Fluctuation and Noise Letters*, 4:L107–L118.

Cabrera, J. and Milton, J. (2004). Human stick balancing : Tuning levy flights to improve balance control. *Physical Review Letters*, 14:691–698.

Dragan, V. and Morozan, T. (1999). Stability and robust stabilization to linear stochastic systems described by differential equations with markovian jumping and multiplicative white noise. *The Institute of Mathematics of the Romanian Academy*, Preprint 17/1999.

Dragan, V. and Morozan, T. (2004). The linear quadratic optimiziation problems for a class of linear stochastic systems with multiplicative white noise and markovian jumping. *IEEE Transactions on Automat. Contr.*, 49:665–675.

Fen, X., Loparo, K., Ji, Y., and Chizeck, H. (1992). Stochastic stability properties of jump linear systems. *IEEE Transactions on Automat. Contr.*, 37:38–53.

Hu, S., Liao, X., and Mao, X. (2003). Stochastic hopfield neural networks. *Journal of Physics A:Mathematical and General*, 36:1–15.

Kaufman, H., Barkana, I., and Sobel, K. (1998). *Direct Adaptive Control Algorithms - Theory and Applications*. Springer (New-York), 2 edition.

Kwok, H., Zhong, G., and Tang, W. (2003). Use of neurons in chaos generation. In *ICONS 2003, Portugal*.

Löfberg, J. (2004a). *YALMIP : A Toolbox for Modeling and Optimization in MATLAB*.

Löfberg, J. (2004b). YALMIP : A toolbox for modeling and optimization in MATLAB. In *Proceedings of the CACSD Conference*, Taipei, Taiwan.

Lure, A. and Postnikov, V. (1944). On the theory of stability of control systems. *Applied Mathematics and Mechanics (in Russian)*, 8(3):245–251.

Mazenc, F. and Niculescu, S.-I. (2001). Lyapunov stability analysis for nonlinear delay system. *Systems and Control Letters*, 42:245–251.

Poznyak, A. and Sanchez, W. Y. E. (1999). Identification and control of unknown chaotic systems via dynamic neural networks. *IEEE Transactions on Circuits and Systems E: Fundamental Theory and Applications*, 46:1491–1495.

Stoica, A. and Yaesh, I. (2005). Hopfield networks with jump markov parameters. *WSEAS Transactions on Systems*, 4:301–307.

Stoica, A. and Yaesh, I. (2006). Delayed hopfield networks with multiplicative noise and jump markov parameters. In *MTNS 2006, Kyoto*.

X. Liao, G. C. and Sanchez, E. (2002). Lmi based approach to asymptotically stability analysis of delayed neural networks. *IEEE Transactions on Circuits and Systems E: Fundamental Theory and Applications*, 49:1033–1039.

Yaesh, I. and Shaked, U. (2005). Stochastic passivity and its application in adaptive control. In *CDC 2005, Kyoto*.

# DISCRETE-TIME ADAPTIVE REPETITIVE CONTROL
## *Internal Model Approach*

Andrzej Krolikowski and Dariusz Horla
*Poznan University of Technology, Institute of Control and Information Engineering*
*Division of Control and Robotics, ul. Piotrowo 3a, 60-965, Poland*
*Andrzej.Krolikowskia@put.poznan.pl*

Keywords: Discrete-time systems, IMC structure, Adaptive repetitive control.

Abstract: Repetitive control is known as one of the most effective methods to reduce repetitive errors with a known period in various practical control systems performing repetitive tasks. The application of Internal Model Control (IMC) structure for repetitive control is introduced. Two IMC-based repetitive control configurations are proposed together with their adaptive versions. A comparative simulation study is carried out for the model of a first link of the robot.

## 1 INTRODUCTION

Many computer-controlled control systems perform repetitive (periodic) tasks thus being subjected to repetitive as well as nonrepetitive disturbances. Rejecting of periodic disturbances or tracking a periodic reference signal can be considered as the original aim of the repetitive controller. In last years much effort has been devoted to the development of discrete-time repetitive control systems which may be considered to be very powerful tools to regulate the repetitive errors whose fundamental frequencies are priori known (Hillerström and Walgama, 1996; Chang et al., 1995; Kempf et al., 1993; Hu and Yu, 1996). The case of uncertain period time is analyzed in (Steinbuch, 2002). Usually, the repetitive errors containing only one fundamental frequency and its harmonics are taken for consideration. A discrete-time repetitive controller for odd harmonic reference and disturbance signals is proposed in (Griñó and Costa-Castelló, 2005). This type of signals appear for example in power electronics systems. Usually, the period of repetitive signals is assumed to be known. In (Steinbuch, 2002), a new structure for repetitive control is proposed which is robust for changes in period-time. The problem of tracking arbitrary periodic reference signals is discussed in (Ledwich and Bolton, 1993), where the compensator design is proposed to give zero steady-state error. The robustness issues of repetitive control are for example examined in (Chang et al., 1995; Hu and Yu, 1996; Tenney and Tomizuka, 1996). The problem of adaptive repetitive control is not much discussed in the literature.

In this paper, two structures of the adaptive repetitive IMC system are presented and simulated using the model of one link of the robot.

## 2 THE INTERNAL MODEL PRINCIPLE

A block diagram of the conventional discrete-time repetitive control system based on the Internal Model Principle (IMP) for a single fundamental frequency of repetitive errors is shown in Fig.1.

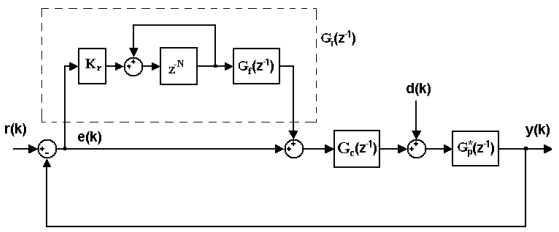



Figure 1: IMP-based repetitive control system.

In this block diagram $r(k)$ and $d(k)$ represent the unknown periodic reference and disturbance with known period, respectively. Typically, the disturbance is assumed to have one fundamental frequency $f_o$ and higher harmonics. The gain $K_r$ is an adjustable parameter of the repetitive controller $G_r(z^{-1})$.

The IMP implies a use of the repetitive signal generator which is a $N$ step delay chain with positive feedback around it (Hillerström and Walgama, 1996) having the transfer function

$$G_{im}(z^{-1}) = \frac{z^{-N}}{1-z^{-N}} \tag{1}$$

This generator represents simply the model of a periodic disturbance. If $T_s$ denotes the sampling period then $NT_s$ is chosen to be equal to the period of the fundamental component of the repetitive errors, i.e. $NT_s = T_o = \frac{1}{f_o}$ so $N = \frac{T_o}{T_s}$. A harmonic signal has only one component at $\frac{2\Pi k}{NT_s}$ rad s$^{-1}$ for $k = 1, 2, \cdots$.
Let the plant be given by the transfer function $G_p^*(z^{-1})$. It is known (Kempf et al., 1993) that for the repetitive control system design a parametric model of the plant is required. The nominal plant is characterized by the transfer function

$$G_p(z^{-1}) = z^{-d}\frac{B(z^{-1})}{A(z^{-1})} \tag{2}$$

with $B(z^{-1}) = b_1 z^{-1} + \cdots + b_{nb} z^{-nb}$, $A(z^{-1}) = 1 + a_1 z^{-1} + \cdots + a_{na} z^{-na}$ and $d \geq 0$.
A nominal feedback controller $G_c(z^{-1})$, typically a lag-lead compensator or PD controller is designed so that for the nominal open-loop transfer function $G_o(z^{-1}) = G_c(z^{-1})G_p(z^{-1})$, the nominal closed-loop transfer function

$$G(z^{-1}) = \frac{G_o(z^{-1})}{1+G_o(z^{-1})} \tag{3}$$

is asymptotically stable and minimumphase. To assure the stability of the control system with repetitive controller the filter $G_f(z^{-1})$ such that

$$G_f(z^{-1})G(z^{-1}) = 1 \tag{4}$$

is usually introduced (Chang et al., 1995; Kempf et al., 1993; Chang et al., 1998).

# 3 THE MULTIPLE REPETITIVE CONTROL SYSTEM

The purpose of the multiple repetitive controller is to regulate multiple repetitive errors which contain multiple dominant fundamental frequencies and their harmonics (Chang et al., 1998). The multiple repetitive discrete-time control system is depicted in Fig.2. It is worthy to note that all repetitive control systems can be augmented by multiple repetitive loops.

Consider again the unmodelled dynamics in the form of a multiplicative modelling uncertainty given by $G^*(z^{-1}) = G(z^{-1})[1+\Delta(z^{-1})]$. Then from (4), a relationship between $G_f(z^{-1})$ and $G^*(z^{-1})$ can be obtained in terms of modelling uncertainty

$$G_f(z^{-1})G^*(z^{-1}) = 1 + \Delta(z^{-1})] \tag{5}$$

From (3),(4) and (5), a modelling uncertainty can be derived as

$$\Delta(z^{-1}) = \frac{G_p^*(z^{-1}) - G_p(z^{-1})}{G_p(z^{-1})(1+G_o^*(z^{-1}))} \tag{6}$$

where $G_o^*(z^{-1}) = G_c(z^{-1})G_p^*(z^{-1})$.

Assuming that $|\Delta(z^{-1})| \leq \varepsilon$ for each $z$ such that $|z| \geq 1$, the robust stability can be demonstrated (Chang et al., 1998) provided that the gains $K_{ri}$ satisfy the condition

$$\sum_{i=1}^{n} K_{ri} < \frac{2}{1+\varepsilon}. \tag{7}$$

# 4 THE INTERNAL MODEL CONTROL STRUCTURE

## 4.1 The Main IMC Configuration

The discrete-time IMC (Internal Model Control) system structure is shown in Fig.3. This structure is a counterpart of the continuous-time IMC controller given in (Datta, 1998). It is known that every stabilizing controller $G_c(z^{-1})$ is given by

$$G_c(z^{-1}) = \frac{Q(z^{-1})}{1-G_p(z^{-1})Q(z^{-1})} \tag{8}$$

where $Q(z^{-1})$ varies over the set of all stable rational transfer functions. This structure may also yield a stable closed-loop performance for unstable plant

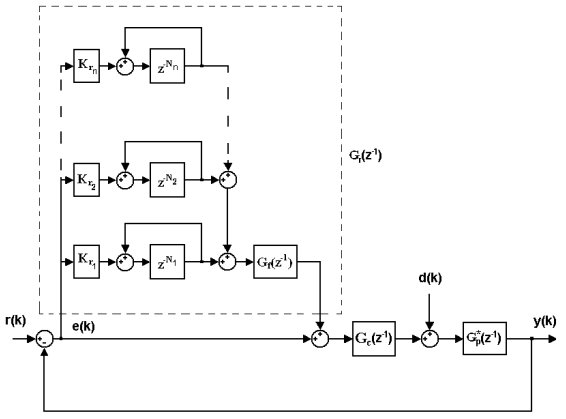



Figure 2: Multiple repetitive control system.

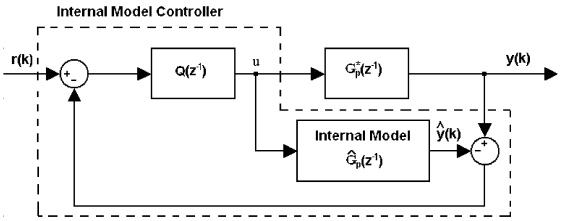



Figure 3: IMC structure.

provided that a plant model $\hat{G}_p(z^{-1})$ is stable, however in this case $Q(z^{-1})$ must not only be stable but must also satisfy certain constraints imposed by unstable poles of the plant.

Suppose that the (possibly proper) $\hat{G}_p(z^{-1}), Q(z^{-1})$ are stable so that the IMC structure is stable for $\hat{G}_p(z^{-1}) = G_p^*(z^{-1})$. Let the uncertainty modelling have the following multiplicative form

$$G_p^*(z^{-1}) = \hat{G}_p(z^{-1})[1+\Delta_p(z^{-1})] \tag{9}$$

where $\Delta_p(z^{-1})$ is stable strictly proper uncertainty. From the IMC structure (Fig.3) the following equation can be derived

$$u(k) = -\hat{G}_p(z^{-1})Q(z^{-1})\Delta_p(z^{-1})u(k) + Q(z^{-1})r(k) \tag{10}$$

so

$$||u(k)||_2 \le ||\hat{G}_p(z^{-1})Q(z^{-1})\Delta_p(z^{-1})||_\infty ||u(k)||_2 + \\ + ||Q(z^{-1})||_\infty ||r(k)||_2 \tag{11}$$

This shows that if

$$||\hat{G}_p(z^{-1})Q(z^{-1})\Delta_p(z^{-1})||_\infty < 1 \tag{12}$$

then

$$||u(k)||_2 \le [1-||\hat{G}_p(z^{-1})Q(z^{-1})\Delta_p(z^{-1})||_\infty]^{-1} \times \\ \times ||Q(z^{-1})||_\infty ||r(k)||_2 \tag{13}$$

so the condition (12) gives the sufficient condition for $L_2$ stability, thus the IMC structure is robust with respect to modelling errors in the plant. Note that the closed-loop transfer function is

$$\frac{y(z)}{r(z)} = \frac{\hat{G}_p(z^{-1})Q(z^{-1})\Delta_p(z^{-1}) + \hat{G}_p(z^{-1})Q(z^{-1})}{1+\hat{G}_p(z^{-1})Q(z^{-1})\Delta_p(z^{-1})} \tag{14}$$

For similar approach in continuous-time IMC structure see (Datta, 1998).

## 4.2 The Pole-placement IMC Configuration

The standard RST controller has a form

$$R(z^{-1})u(k) = -S(z^{-1})y(k) + T(z^{-1})r(k+d+1) \tag{15}$$

and is the solution of

$$A(z^{-1})R(z^{-1}) + z^{-d}B(z^{-1})S(z^{-1}) = A(z^{-1})P(z^{-1}) \tag{16}$$

where $P(z^{-1})$ is the stable polynomial the roots of which are assumed to be the closed-loop poles. The above equation implies that

$$S(z^{-1}) = A(z^{-1})S'(z^{-1}), \tag{17}$$

i.e.(16) is replaced by

$$R(z^{-1}) + z^{-d}B(z^{-1})S'(z^{-1}) = P(z^{-1}) \tag{18}$$

and this allows the controller to be characterized by

$$R(z^{-1}) = P(z^{-1}) - z^{-d}B(z^{-1})S'(z^{-1}). \tag{19}$$

Polynomial $S'(z^{-1})$ is assumed to be stable. For example, if $R(z^{-1})$ contains an integrator then

$$S'(1) = \frac{P(1)}{B(1)} \tag{20}$$

yielding

$$R(z^{-1}) = P(z^{-1}) - z^{-d}\frac{B(z^{-1})P(1)}{B(1)} \tag{21}$$

and

$$S(z^{-1}) = A(z^{-1})\frac{P(1)}{B(1)} \tag{22}$$

Using the controller equation (15) and (18) one obtains

$$\frac{P(z^{-1})B(1)}{A(z^{-1})P(1)}u(k) = -[y(k) - z^{-d}\frac{B(z^{-1})}{A(z^{-1})}u(k)] + \\ + \frac{T(z^{-1})B(1)}{A(z^{-1})P(1)}r(k+d+1) \tag{23}$$

which is the IMC scheme as shown in Fig.4 where

$$G_T(z^{-1}) = \frac{T(z^{-1})B(1)}{A(z^{-1})P(1)}. \tag{24}$$

and using the notation from Fig.3

$$Q(z^{-1}) = \frac{A(z^{-1})P(1)}{P(z^{-1})B(1)}. \tag{25}$$

It is easy to see that taking $T(z^{-1}) = \frac{P(1)A(1)}{B(1)}$ guarantees the zero steady-state error in the case of perfect matching.

## 4.3 The Repetitive IMC Configuration

The proposed repetitive IMC system structure is represented in Fig.5. This is a combination of the IMC structure (Figs.3,4) and the standard repetitive controller (or multiple repetitive controller). The aim of

this control system is reject the repetitive errors by the repetitive controller and to improve the robustness by a proper choice of $Q(z^{-1}$.

From (3), (4), (9) and (13) the following relation between uncertainties $\Delta_p(z^{-1})$ and $\Delta(z^{-1})$ can be found

$$\Delta(z^{-1}) = \frac{\Delta_p(z^{-1})}{1+G_o^*(z^{-1})}. \qquad (26)$$

Taking into account that $|\Delta(z^{-1})| \leq \varepsilon$ as in (6) the following condition can be derived

$$|\frac{\Delta_\mu(z^{-1})}{1+G_o^*(z^{-1}))}| \leq \varepsilon \qquad (27)$$

This means that under this condition the robust stability of the repetitive IMC structure will be assured if additionally the uncertainty $\Delta(z^{-1})$ is stable. The inequality (27) can not practically be checked out because $G_o^*(z^{-1})$ is not known, however using (8) and (9) the inequality $|\Delta(z^{-1})| \leq \varepsilon$ takes a form

$$|\frac{\Delta_p(z^{-1})(1-G_p(z^{-1})Q(z^{-1}))}{1+G_p(z^{-1})Q(z^{-1})\Delta_p(z^{-1})}| \leq \varepsilon \qquad (28)$$

so the (multiple) repetitive IMC system is robustly stable if the uncertainty $\Delta_p(z^{-1})$ is such that the above condition is fulfilled.

### 4.4 The Adaptive Repetitive IMC Structure

The proposed adaptive repetitive IMC system structure is represented in Fig.6, where the parameter estimation is realized using the standard recursive least-squares algorithm. The adaptation is realized in an

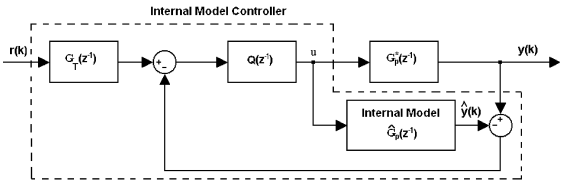



Figure 4: Pole-placement IMC structure.

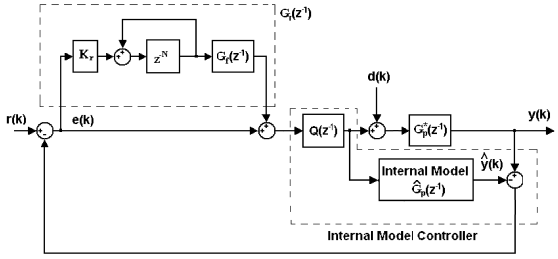



Figure 5: Repetitive IMC structure.

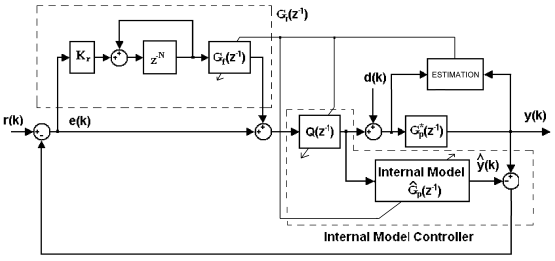



Figure 6: Adaptive repetitive IMC structure.

indirect way, i.e. the model parameters are first estimated, and subsequently the obtained parameter estimates $\hat{\theta}(k) = (\hat{a}_1(k), \ldots, \hat{a}_{na}(k), \hat{b}_1(k), \ldots, \hat{b}_{nb}(k))^T$ are used for tuning the parameters of both repetitive and internal model controllers.

## 5 SIMULATIONS

Often robotic manipulators are required to execute repetitive tasks. Then the desired trajectory to be followed by the manipulator is bounded and periodic with known period. Below a first link of the AdeptOne robot (Tenney and Tomizuka, 1996) is taken as an example for simulations. The link considered as a plant is approximated by the nominal ARX model

$$G_p(z^{-1}) = \frac{0.000242z^{-1}}{1-1.9788z^{-1}+0.9789z^{-2}} \qquad (29)$$

obtained at $\frac{1}{T_s} = 1kHz$ sampling rate. The nominal compensator has a form of PD-type

$$G_c(z^{-1}) = 119.5\frac{1-0.925z^{-1}}{1-0.65z^{-1}}. \qquad (30)$$

The main IMC repetitive controller has been tested for

$$Q(z^{-1}) = \frac{119.5-347z^{-1}+335.7z^{-2}-108.2z^{-3}}{1-2.6z^{-1}+2.238z^{-2}-0.6363z^{-3}} \qquad (31)$$

that has been obtained according to (8) for a stable plant model (29). In turn, the filter $G_f(z^{-1})$ was derived according to (4) as

$$G_f(z^{-1}) = \frac{1-4.55z^{-1}+8.278z^{-2}-7.529z^{-3}}{0.02892z^{-1}-0.08397z^{-2}+0.08124z^{-3}-0.02619z^{-4}} + \frac{3.424z^{-4}-0.6229z^{-5}}{0.02892z^{-1}-0.08397z^{-2}+0.08124z^{-3}-0.02619z^{-4}}. \qquad (32)$$

The disturbance $d(k)$ with amplitude of 5 units contains the fundamental and harmonic frequencies of $f_{o1} = 5Hz$ $(10Hz, 15Hz)$, $f_{o2} = 7Hz$ $(14Hz, 21Hz)$, $f_{o3} = 9Hz$ $(18Hz, 27Hz)$ thus $N_1 = 200, N_2 = 143$,

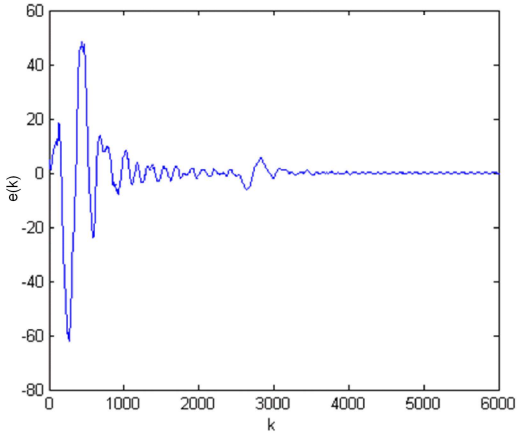



Figure 7: Adaptive repetitive IMC, disturbance attenuation.

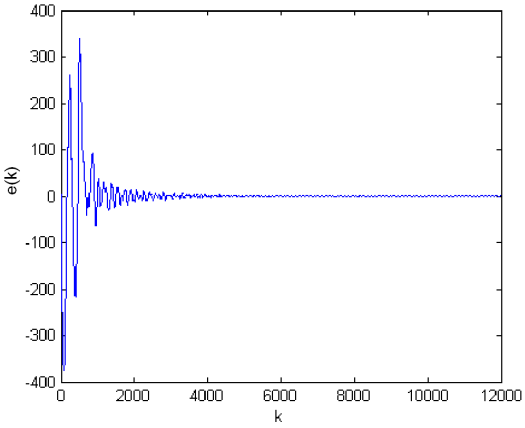



Figure 9: Adaptive repetitive pole-placement IMC.

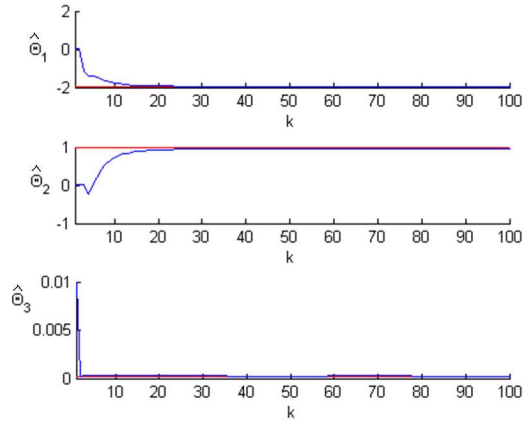



Figure 8: Adaptive repetitive IMC, parameter estimates.

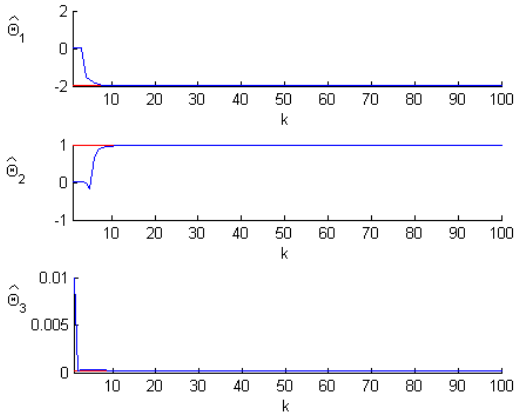



Figure 10: Adaptive repetitive pole-placement IMC.

$N_3 = 111$ with $K_{r1} = K_{r2} = K_{r3} = 0.5$. Additionally, a pulse disturbance $d_p$ with amplitude of 15 units is also inserted to the input of the plant.

The initial conditions for parameter estimates and covariance matrix in the recursive least squares algorithm were taken as $\hat{\theta}(0) = (0.01, 0.01, 0.01)^T$ and $P(0) = 100I$.

The performance of adaptive multiple repetitive IMC control system given in Fig.7 shows the effect of disturbance attenuation. The corresponding parameter estimates are shown in Fig.8.

Finally, the adaptive pole-placement IMC structure was combined with multiple repetitive controller. For the polynomial $P(z^{-1}) = 1 - 1.8z^{-1} + 0.9z^{-2}$ one obtains from (25)

$$Q(z^{-1}) = \frac{0.1 - 0.1979z^{-1} + 0.09789z^{-2}}{0.000242 - 0.0004356z^{-1} + 0.0002178z^{-2}}, \quad (33)$$

and from (24)

$$G_T(z^{-1}) = \frac{0.0001}{1 - 1.979z^{-1} + 0.9789z^{-2}}. \quad (34)$$

The filter $G_f(z^{-1})$ was derived again from (4), however in this case the transfer function $G(z^{-1})$ is

$$G(z^{-1}) = \frac{G_T(z^{-1})Q(z^{-1})G_p(z^{-1})}{1 + G_T(z^{-1})Q(z^{-1})G_p(z^{-1})}. \quad (35)$$

The error signal is shown in Fig.9, and the corresponding parameter estimates are shown in Fig.10 for multiple harmonic disturbance attenuation.

## 6 CONCLUSIONS

Two structures of IMC repetitive control system are examined and their adaptive versions are simulated taking the first link of an AdeptOne robot as the example. The proposed control structures can be enlarged by the multiple repetitive controller. The adaptive loop included into the IMC repetitive control system reduces the level of parametric uncertainty thus improves the quality of disturbance attenuation. In

this way the proposed configurations can be considered as the robust adaptive ones.

## REFERENCES

Chang, W., Suh, I., and Kim, T. (1995). Analysis and design of two types of digital repetitive control systems. *Automatica*, 31(5):741–746.

Chang, W., Suh, I., and Oh, J.-H. (1998). Synthesis and analysis of digital multiple repetitive control systems. In *Proceedings of the ACC*, pages 2687–2691, Philadelphia.

Datta, A. (1998). Adaptive internal model control. *Springer*.

Griñó, R. and Costa-Castelló, R. (2005). Digital repetitive plug-in controller for odd-harmonic periodic references and disturbances. *Automatica*, 41:153–157.

Hillerström, G. and Walgama, K. (1996). Repetitive control theory and applications - a survey. In *Proceedings of the 13th IFAC World Congress*, pages CD–ROM, San Francisco.

Hu, J. and Yu, S.-H. (1996). Optimal repetitive control system design using mixed time and frequency domain criteria. In *Proceedings of the 13th IFAC World Congress*, pages CD–ROM, San Francisco.

Kempf, C., Messner, W., Tomizuka, M., and Horowitz, R. (1993). Comparison of four discrete-time repetitive control algorithms. *IEEE Control Systems*, 13(6):48–54.

Ledwich, G. and Bolton, A. (1993). Repetitive and periodic controller design. *IEEE Proceedings-D*, 140(1):19–24.

Steinbuch, M. (2002). Repetitive control for systems with uncertain period time. *Automatica*, 38:2103–2109.

Tenney, J. and Tomizuka, M. (1996). Effects of non-periodic disturbances on repetitive control systems. In *Proceedings of the 13th IFAC World Congress*, pages CD–ROM, San Francisco.

# OFF-LINE ROBUSTIFICATION OF EXPLICIT MPC LAWS
## *The Case of Polynomial Model Representation*

Pedro Rodríguez-Ayerbe and Sorin Olaru
*Department of Automatic Control, Supélec, 3 rue Joliot Curie, F91192 Gif-sur-Yvette, France*
*pedro.rodriguez@supelec.fr, sorin.olaru@supelec.fr*

Keywords: Piecewise affine controller, Robustification, Youla-Kucera parameter, Model Predictive Control.

Abstract: The paper deals with the predictive control for linear systems subject to constraints, technique which leads to nonlinear (piecewise affine) control laws. The main goal is to reduce the sensitivity of these schemes with respect to the model uncertainties and avoid in the same time a fastidious on-line optimisation which may reduce the range of application. In this idea a two stage predictive strategy is proposed, which synthesizes in a first instant an analytical (continuous and piecewise linear) control law based on the nominal model and secondly robustify the central controller (the controller obtained when no constraint is active). This robustification is then expanded to all the space of the piecewise structure by means of its corresponding noise model.

# 1 INTRODUCTION

The model predictive control (MPC) laws are optimization based techniques which allow constraints handling from the design stage. The analytical formulation of the optimum and its on-line evaluation avoids a challenging optimization from the point of view of the real-time control environment. Solutions in this direction exist at least for two important classes of problems (linear and quadratic) subject to linear constraints due to the Abadie constraint qualification (Goodwin *et al*., 2004). It must be said that these are in fact a part of a larger class of multiparametric convex programs (Bemporad *et al*., 2002b) for which exact or approximate algorithms exist (TØndel *et al*., 2003, Seron *et al*, 2003, Olaru and Dumur, 2004; Bemporad and Filippi, 2006).

In the case of robust predictive control laws, the model uncertainties and the disturbances can be taken into account at the design stage. A popular technique in this sense is the use of a min-max criterium (in the case when the extreme combination of disturbances or uncertainties are known) (Kerrigan and Maciejowski, 2004; Bemporad *et al*., 2002a) which comes finally to the resolution of a single multiparametric linear program. The structure of this ultimate optimization is however quite complex and large prediction horizons cannot be handled due to the exponential growth of disturbances realization to be taken into account. In a slightly different manner, by constructing an estimation mechanism (Goodwin *et al*., 2004) for the constrained variables, one can obtain alternatively a robust control structure, but the multiparametric optimization remains intricate.

A first study on the robustness improvements for the explicit affine feedback policy constructed upon constrained predictive control strategies was presented in (Olaru and Rodriguez-Ayerbe, 2006). The simplest way to proceed is to consider an observer of the state variables (Goodwin *et al*., 2004), the dimension of the state space being preserved and the piece-wise structure of controller unchanged. The same observer can be used for all feasible regions and can be viewed as noise characterisation of the model. Nevertheless, the observer does not describe the entire class of stabilizing controllers. The present paper presents an improved result based on the Youla-Kučera parametrization which spans the space of stabilizing controllers. For a two-degree of freedom controller, one has access to all the stabilizing controllers that preserve the same input/output behavior, so the Youla-Kučera parameter offers more degrees of freedom than the use of an observer.

The robustification is made such that the state space dimension of the controller is augmented. The main contribution here is the reconstruction of the noise model induced by the central Youla-Kučera

parameter, in order to use it to generate the corresponding robust piece-wise controller.

In the following, section 2 briefly recalls the predictive control and the Youla-Kučera parametrization. Section 3 details the explicit formulation of the control laws obtained in the constrained case. Section 4 contains the main contribution: the noise model of the Youla-Kučera parameter and the numerical examples are presented in section 5 and the final conclusions in section 6.

## 2 PREDICTIVE CONTROL

The Generalized Predictive Control (GPC) strategy, introduced in (Clarke *et al*., 1987), uses for the prediction a CARIMA plant model:

$$A(q^{-1})y_t = B(q^{-1})u_{t-1} + \frac{C(q^{-1})\xi_t}{\Delta(q^{-1})} \tag{1}$$

with $u$, $y$ the input and output, $\xi$ a white noise, $A$ and $B$ polynomials in the backward shift operator of degrees $n_a$ and $n_b$ respectively, and $\Delta(q^{-1}) = 1 - q^{-1}$ the difference operator. The $C$ polynomial is the model argument taking into account the noise influence on the system. In the GPC case the cost function to be minimized over a receding horizon is quadratic:

$$J = \sum_{j=N_1}^{N_2} \left[ w_{t+j} - \hat{y}_{t+j} \right]^2 + \sum_{j=1}^{N_u} \lambda_j \left[ \Delta u_{t+j-1} \right]^2 \tag{2}$$

where $N_1, N_2$ are the costing horizons, $N_u$ the control horizon, $\lambda_j$ the control weighting factor and $w$ the set-point.

Using the model (1) and the solution of some Diophantine equations (Clarke *et al*., 1987), this control strategy leads to two-degrees of freedom *RST* controller, implemented through a difference equation (Figure 1):

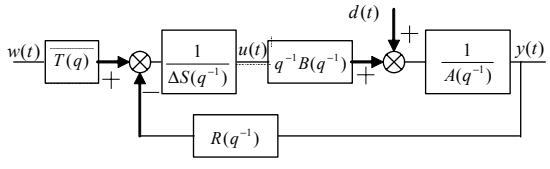



Figure 1: Two-degrees of freedom GPC controller.

In (Yoon and Clarke, 1995) the relation between the RST controller obtained with $C$=1 and $C \neq 1$ is studied. Considering $R', S', T'$ the controller obtained with $C$=1 and $\overline{R}, \overline{S}, \overline{T}$ whose obtained with $C \neq 1$, the following relations are obtained:

$$\overline{R} = R'C + A\Delta M \quad \overline{S} = S'C - q^{-1}BM \quad \overline{T} = T'C \tag{3}$$

with:

$$M(q^{-1}) = \sum_{i=N_1}^{N_2} \alpha_i q^i (CE_i' - \overline{E}_i) \tag{4}$$

$$\mathbf{m} = \begin{bmatrix} \alpha_{N1} & \dots & \alpha_{N2} \end{bmatrix} \tag{5}$$

$\mathbf{m}$ being the first row of $(\mathbf{G}^{\mathrm{T}}\mathbf{G} + \mathbf{\Lambda})^{-1}\mathbf{G}^{\mathrm{T}}$.

The set of all stabilizing controllers for the system shown in Figure 1 is given by the Youla-Kučera parametrization as follows (Maciejowski, 1989):

$$Q = \frac{Q_{num}}{Q_{den}} \begin{cases} R = R'Q_{den} + \Delta A Q_{num} \\ S = S'Q_{den} - q^{-1}BQ_{num} \\ T = T'Q_{den} \end{cases} \tag{6}$$

where $Q(q^{-1})$ is a stable transfer function.

The choice of the $Q$ parameter is a complex problem on its own but it is not the subject of the current paper. The methods presented in (Rodriguez and Dumur, 2005; Rossiter 2003; Ansay *et al*., 1998; Yoon and Clarke, 1995; Kouvaritakis *et al*., 1992) can be used for the choice of this parameter.

Comparing (3) and (6) it turns out that the controller for $C \neq 1$ is obtained for $Q$=$M$/$C$. As $M$ depends of $C$ as shown by (4), the robustification by the $C$ polynomial has less degrees of freedom than the robustification by Youla-Kučera parameter (Yoon and Clarke, 1995).

## 3 EXPLICIT CONSTRAINED GPC LAWS

In the case when the GPC law is subject to constraints, the optimization has to be solved with respect to a feasible domain. If the considered constraints are stated on the control action, on the control increment, on the plant outputs or any other signal related by a CARIMA model to the control signal, then one can restate them in a form depending only on the control increment, leading to a set of linear constraints (Ehrlinger *et al*., 1996):

$$\begin{cases} \mathbf{A}_{in}\ \mathbf{k}_u \le \mathbf{B}_{in}\ \mathbf{p}_t + \mathbf{b}_{in} \\ \mathbf{A}_{eq}\ \mathbf{k}_u = \mathbf{B}_{eq}\ \mathbf{p}_t + \mathbf{b}_{eq} \end{cases} \quad (7)$$

involving the optimization argument $\mathbf{k}_u(t)$ and the vector of context parameters:

$$\mathbf{p}_t = \left[\mathbf{y}_{past} \quad \{\boldsymbol{\gamma}_{past}\}^{\mathrm{T}} \quad \boldsymbol{\Delta}\mathbf{u}_{past}^{\mathrm{T}} \quad \mathbf{w}^{\mathrm{T}}\right]^{\mathrm{T}} \quad (8)$$

which regroups a finite sequence of past inputs $\boldsymbol{\Delta}\mathbf{u}$, future setpoints $\mathbf{w}$, past outputs $\mathbf{y}$ and present and future values of the signals under constraints (noted for short $\gamma_{past}$).

It is interesting to observe that this set of constraints characterizes in fact a parameterized polyhedron (Olaru and Dumur, 2004) in the optimization argument space. The optimum will lie on a combination of its parameterized vertices and thus one can construct the explicit solution for the multiparametric optimization:

The use of a dual representation of the feasible domain and projection mechanisms (see Olaru and Dumur, 2004 and 2005) can be advantageous in order to express the optimality conditions if there exist unbounded feasible directions.

$$\min_{\mathbf{k}_u(t)} J$$
$$\text{subject to}: \begin{cases} \mathbf{A}_{in}\mathbf{k}_u(t) \le \mathbf{B}_{in}\mathbf{p}_t + \mathbf{b}_{in} \\ \mathbf{A}_{eq}\mathbf{k}_u(t) = \mathbf{B}_{eq}\mathbf{p}_t + \mathbf{b}_{eq} \end{cases} \quad (9)$$

Subsequently, the predictive control law can be described explicitly upon the solution of (9) as a piecewise affine function of the vector of parameters (Seron *et al*, 2003; Bemporad *et al*., 2002; Olaru and Dumur, 2004).

$$\mathbf{k}_u^*(\mathbf{p}) = K_{Lin_i}\mathbf{p} + K_{tli} \quad \text{for} \quad \mathbf{p} \in CR_i \quad (10)$$

with $CR_i$, critical regions in the space of context parameters, The GPC subjet to constraints has a piecewise RST polynomial form:

$$S_i(q^{-1})\Delta u_t = -R_i(q^{-1})y_t + T_i(q)w_t + V_i\gamma_{past} + tl_i \quad (11)$$

for $\left[\mathbf{y}_{past} \quad \{\boldsymbol{\gamma}_{past}\}^{\mathrm{T}} \quad \boldsymbol{\Delta}\mathbf{u}_{past}^{\mathrm{T}} \quad \mathbf{w}^{\mathrm{T}}\right]^{\mathrm{T}} \in CR_i$ .

The structure of such a piecewise controller is shown in Figure 2. Once the look-up table of local polynomial RST laws is available, an efficient positioning mechanism (based on a search tree) can be constructed such that the on-line evaluation routine can find the optimal control action according to the GPC philosophy (TØndel *et al*., 2003).

# 4 ROBUSTIFICATION

At this stage, it is assumed that the design of initial controller has been performed with $N_1$, $N_2$, $N_u$, $\lambda$ adjusted to satisfy the required input/output behavior. The resulting picewise two-degrees of freedom *RST* controller will be denoted $R_i', S_i', T_i'$, in the following sections.

The observer based robustification corresponds to the consideration of an observer of the state. In the case of GPC, this corresponds to a choice of a *C* polynomial. The roots of this polynomial correspond to the poles of the observer. The obtained piecewise controller can be implemented as in Figure 2, see also (Olaru and Rodriguez-Ayerbe, 2006, Camacho and Bordons, 2004 ; Bitmead *et al.*,1990).

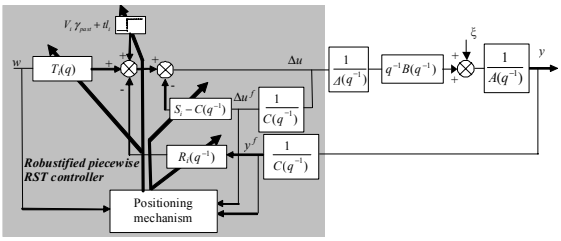



Figure 2: Piecewise RST formulation for the GPC law under constraints and robustification using the C polynomial.

The Youla-Kučera parameter based robustification has more degrees of freedom than the use of an observer. It allows to access the entire space of stabilizing controllers. The idea is thus to robustify the central RST law of the piecewise controller, that is, the law where the constraints are not activated, and to expand this robustification to the others RST laws of the initial piecewise controller. The choice of this parameter can be done using stability robustness and nominal performance specifications, see (Rodriguez and Dumur, 2005; Rossiter 2003; Ansay *et al*., 1998; Yoon and Clarke, 1995; Kouvaritakis *et al.*, 1992). In the following the selection will be done according to these principles.

As the use of an observer corresponds to the consideration of a *C* polynomial for the noise influence on the CARIMA model (1), the idea is to find the corresponding noise model of the Youla-Kučera parameter. To do this, an extra polynomial is added in the model of the system. A $D(q^{-1})$ polynomial will appear as following:

$$A(q^{-1})y_t = B(q^{-1})u_{t-1} + \frac{C(q^{-1})\xi_t}{D(q^{-1})\Delta(q^{-1})} \quad (12)$$

With this new model, Diophantine equations are:

$$\begin{aligned}&\Delta(q^{-1})A(q^{-1})D(q^{-1})E_j(q^{-1})+q^{-j}F_j(q^{-1})=C(q^{-1})\\&G_j(q^{-1})C(q^{-1})+q^{-j}H_j(q^{-1})=B(q^{-1})E_j(q^{-1})D(q^{-1})\end{aligned} \quad (13)$$

Finding the relation between the controller obtained for $C$=$D$=1 and the one obtained for $C \neq 1$ and $D \neq 1$, we obtain something similar to (3). Considering $R', S', T'$ the controller obtained with $C$=$D$=1 and $\widetilde{R}, \widetilde{S}, \widetilde{T}$ whose obtained with $C \neq 1$ and $D \neq 1$, the following relations are obtained:

$$\widetilde{R} = R'C + A\Delta\widetilde{M} \quad \widetilde{S} = S'C - q^{-1}B\widetilde{M} \quad \widetilde{T} = T'C \quad (14)$$

With, (see Appendix for structural details) :

$$\widetilde{M}(q^{-1}) = \sum_{i=N_1}^{N_2} \alpha_i q^i (CE_i' - \widetilde{E}_i D) \quad (15)$$

So, the $D$ polynomial corresponding to the considered Youla-Kučera parameter must verify:

$$Q_{num}(q^{-1}) = \widetilde{M}(q^{-1}) = \sum_{i=N_1}^{N_2} \alpha_i q^i (CE_i' - \widetilde{E}_i D) \quad (16)$$

Once the corresponding noise model has been obtained, it can be used to regenerate the piecewise affine controller. The same input/output behaviour as for the initial one is assured, in the ideal case of no model errors. A modified close loop behaviour will be observed with respect to disturbance rejection, robustness, etc.

The resolution of (16) is a non linear problem that can be undertaken with standard optimization methods. Nevertheless, is not always possible to guarantee a *real* solution. The resolution of (16) and its limitations are raising interesting questions, research being currently conducted on this subject. From a practical point of view, any such limit case can be avoided by retuning the initial predictive control parameters or the robustification specification.

# 5 EXAMPLE

Consider the position control of an induction motor, with 1.0724 ms as sampling period

$$H(q^{-1}) = \frac{\theta(q^{-1})}{\tau_{\text{ref}}(q^{-1})} = \frac{10^{-4}(0.821q^{-1} + 0.8206q^{-2})}{(1-q^{-1})(1-0.998q^{-1})} \quad (17)$$

Constraints in control amplitude are considered: $\tau_{\text{ref}} \in [\tau_{\max}, -\tau_{\max}]$ and $\tau_{\max} = 1.8$. An initial GPC controller is designed with $C = D = 1$ with the following tuning parameters: $N_1 = 1$, $N_2 = 16$, $\lambda = 0.0001$ and $N_u = 2$. The position of the motor is obtained through an encoder of 14400 points per rotation, and the highs dynamics of the system (current loop, inverter dynamic, mechanic dynamics in high frequency) have been not identified.

This initial controller is obtained with (9). A piecewise linear controller with 9 regions is obtained. The central region corresponds to the case where no constraint is active. This controller will be noted $R_0, S_0, T_0$.

To robustify off-line this piecewise controller, the idea is to robustify the central one ($R_0, S_0, T_0$) and expand this robustification to other regions. In this way a Youla-Kučera parameter has been obtained by method described in (Rodriguez and Dumur, 2005). The following parameter is considered.

$$Q = \frac{-4196.2 + 10499.99q^{-1} + 8902.17q^{-2} + 2541.93q^{-3}}{1 - 3.565q^{-1} + 4.838q^{-2} - 2.973q^{-3} + 0.7q^{-4}} \quad (18)$$

With (6), we obtain the controller $R_{0Q}, S_{0Q}, T_{0Q}$.

Solving (16) with $C = Q_{den}$, the following D polynomial is obtained:

$$D = 1 - 0.873q^{-1} + 0.472q^{-2} - 0.018q^{-3} + 0.426q^{-4} \quad (19)$$

This value has been obtained by available optimization methods (classical Matlab routines in occurrence) as long as (16) represents a set of non linear equations difficult to solve analytically.

With this $D$ polynomial, the optimization problem (9) can be solved but this time with matrices obtained from (13) for $C = Q_{den}$ and $D$ as in (19). The solution of this new optimization problem leads to a new piecewise controller with 9 regions, as the initial one. The central controller of this piecewise controller correspond to $R_{0Q}, S_{0Q}, T_{0Q}$.

Figures 3 and 4 show the obtained simulations results for a filtered step reference considering a second order neglected dynamic in high frequency of the following characteristics: $\omega_0 = 1000 rad/s$ $\xi = 0.3$.

In these figures we can observe that the obtained behaviour is stable in the case of robustified controller and instable in the case of initial controller. So, the robustified controller has better

behaviour towards uncertainties in high frequency and the continuity between regions is guaranteed.

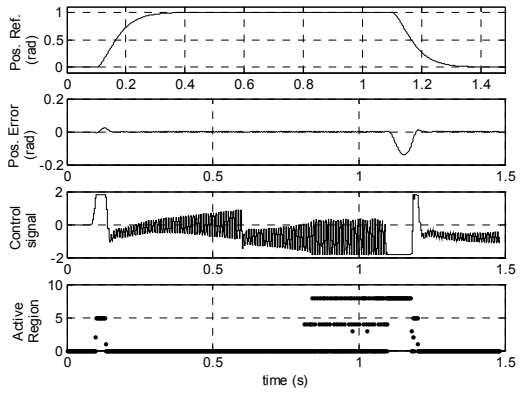



Figure 3: Position reference, position error, control signal and active region for the initial controller and uncertain model.

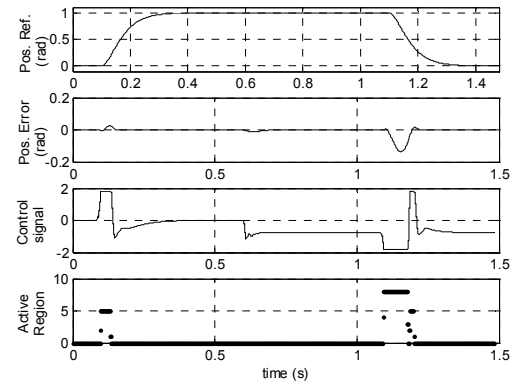



Figure 4: Position reference, position error, control signal and active region for the robustified controller and uncertain model.

# 6 CONCLUSIONS

The paper investigated the robustification methods for the control laws obtained in a constrained predictive control framework. The idea is to design in a first instance a piecewise polynomial controller which satisfy the basic demands in terms of tracking performances. In a second stage, the same predictive control structure (prediction horizon, weightings, etc.) is robustified using the model arguments accounting for the noise influence. The idea is similar to that of using a fixed observer, but exploring all the class of stabilizing controllers of the unconstrained system. This increases the number of degrees of freedom.

The robustification of initial unconstrained controller is made using the Youla-Kučera parametrization, and then this robustification is expanded to all the piecewise structure of the controller. For this, the noise model corresponding to the Youla-Kučera parameter is found, and use to regenerate the robust piecewise controller by preserving the same input/output behavior but being more robust.

The limitations of the method are in the existence of the corresponding noise model of the Youla-Kučera parameter. This is transparent in the resolution of a non linear equation system. The robustification being done off-line, any infeasibility can be handled by retuning the GPC parameters.

# REFERENCES

Ansay P., M. Gevers and V. Wertz (1998). Enhancing the robustness of GPC via a simple choice of the Youla parameter. *European Journal of Control*, 4, 64-70.

Bemporad A., C. Filippi (2006). An Algorithm for Approximate Multiparametric Convex Programming. *Computational Optimization and Applications*, 35, 87-108.

Bemporad A., F. Borelli and M. Morari (2002a). Model predictive control based on linear programming: The explicit solution. *IEEE Transactions on Automatic Control*, 47, 1974–1985.

Bemporad A., M. Morari, V. Dua and E. Pistikopoulos (2002b). The explicit linear quadratic regulator for constrained systems. *Automatica*, 38, 3–20.

Bitmead R.R., M. Gervers and V. Wertz (1990). *Adaptative optimal control. The thinking Man's GPC*. Prentice Hall. Englewood Cliffs, N.J.

Camacho E.F., C. Bordons, "*Model predictive control*", Springer-Verlag, London, 2nd ed., 2004.

Clarke D. W., C. Mohtadi and P.S. Tuffs (1987). Generalized predictive control – Part I and II. *Automatica*, 23(2), 137-160.

Clarke D.W. and R. Scattolini (1991). Constrained receding-horizon predictive control. *IEE Proceedings-D,* 138(4).

Ehrlinger A., P. Boucher and D. Dumur (1996). Unified Approach of Equality and Inequality Constraints in G.P.C. *5th IEEE Conference on Control Applications*.

Goodwin, G.C., M.M. Seron, J.A. De Dona (2004). *Constrained Control and Estimation*, Springer-Verlag, London.

Kerrigan E. and J.M. Maciejowski (2004). Feedback min-max model predictive control using a single linear program: robust stability and the explicit solution. *International Journal of Robust and Nonlinear Control*, 14, 395-413.

Kouvaritakis B, J.A. Rossiter and A.O.T. Chang (1992). Stable generalized predictive control: an algorithm with guaranteed stability. *IEE Proceedings-D*, 139(4), 349-362.

Maciejowski J.M. (1989). *Multivariable feedback design*, Addison-Wesley, England.

Olaru S. and D. Dumur (2004). A parameterized polyhedra approach for explicit constrained predictive control. *43rd IEEE Conference on Decision and Control*, The Bahamas.

Olaru S. and D. Dumur (2005). Avoiding constraints redundancy in predictive control optimization routines. *IEEE Transactions on Automatic Control*, 50(9), 1459–1466.

Olaru S., P. Rodriguez-Ayerbe (2006). Robustification of explicit predictive control laws. *45rd IEEE Conference on Decision and Control,* San Diego.

Rodríguez P. and D. Dumur (2005), Generalized Predictive Control Robustification Under Frequency and Time-Domain Constraints. *IEEE Transactions on Control Systems Technology*, 13(4), 577-587.

Rossiter J.A. (2003). *Model Based Predictive Control. A practical approach*. CRC Press. Florida, USA.

Seron M., G. Goodwin, and J. D. Dona (2003). Characterisation of receding horizon control for constrained linear systems. *Asian Journal of Control*, 5(2), 271–286.

TØndel P., T. Johansen and A. Bemporad (2003). Evaluation of piecewise affine control via binary search tree. *Automatica*, 39 (5),945-950.

Yoon T.W. and D.W. Clarke (1995). Observer design in receiding-horizon predictive control. *Int. Journal of Control*, 61(1), 171-191.

# APPENDIX

By solving the first Diophantine equation of (13) for *C*=*D*=1 and $C \neq 1$, $D \neq 1$, a relationship between the $R'$ polynomial obtained for *C*=*D*=1 and the $\widetilde{R}$ obtained for $C \neq 1$, $D \neq 1$ is obtained:

$$\begin{cases} \Delta(q^{-1})A(q^{-1})E'_j(q^{-1}) + q^{-j}F'_j(q^{-1}) = 1 \\ \Delta(q^{-1})A(q^{-1})D(q^{-1})\widetilde{E}_j(q^{-1}) + q^{-j}\widetilde{F}_j(q^{-1}) = C(q^{-1}) \end{cases}$$

$$\Rightarrow \widetilde{F}_j = CF'_j + \Delta A q^j (CE'_j - D\widetilde{E}_j)$$

$R$ is obtained as: $R = \sum_{i=N_1}^{N_2} \alpha_i F_i$ (see Yoon and Clarke 1995).

For $\widetilde{R}$ :

$$\widetilde{R} = \sum_{i=N_1}^{N_2} \alpha_i \widetilde{F}_i = \sum_{i=N_1}^{N_2} \alpha_i (CF'_i + \Delta A q^i (CE'_i - D\widetilde{E}_i))$$

$$= CR' + A\Delta \sum_{i=N_1}^{N_2} \alpha_i q^i (CE'_i - D\widetilde{E}_i)$$

$$= CR' + A\Delta M$$

With the same development, the corresponding expression for $\widetilde{S}$ of (14) is obtained. Solving (16) comes to the identification of a *D* polynomial corresponding to $Q_{num}(q^{-1})$. It must be noted that the first *i* coefficients of $E_{i+1}$ are the same than those of $E_i$ polynomials. ($E_i$ is a polynomial of degree *i*-1.) With this, the solution of *D* for the numerical example has been obtained solving the following non linear equations:

$$D = 1 + d_1 q^{-1} + d_2 q^{-2} + d_3 q^{-3} + d_4 q^{-4}$$

$$C = 1 + c_1 q^{-1} + c_2 q^{-2} + c_3 q^{-3} + c_4 q^{-4}$$

$$E' = 1 + e'_1 q^{-1} + \ldots + e'_{15} q^{-15}$$

$$\widetilde{E} = 1 + \widetilde{e}_1 q^{-1} + \ldots + \widetilde{e}_{15} q^{-15}$$

$$M = m_0 + m_1 q^{-1} + m_2 q^{-2} + m_3 q^{-3}$$

$$\left\{ \begin{pmatrix} c_1 & c_2 & c_3 & c_4 \\ c_2 & c_3 & c_4 & 0 \\ c_3 & c_4 & 0 & 0 \\ c_4 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & e'_1 & \cdots & & e'_{15} \\ 0 & 1 & \cdots & & e'_{14} \\ 0 & 0 & \ddots & & e'_{13} \\ 0 & 0 & 0 & 1 \cdots & e'_{12} \end{pmatrix} - \right.$$

$$\left. \begin{pmatrix} d_1 & d_2 & d_3 & d_4 \\ d_2 & d_3 & d_4 & 0 \\ d_3 & d_4 & 0 & 0 \\ d_4 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & \widetilde{e}_1 & \cdots & & \widetilde{e}_{15} \\ 0 & 1 & \cdots & & \widetilde{e}_{14} \\ 0 & 0 & \ddots & & \widetilde{e}_{13} \\ 0 & 0 & 0 & 1 \cdots & \widetilde{e}_{12} \end{pmatrix} \right\} \begin{pmatrix} \alpha_1 \\ \vdots \\ \vdots \\ \alpha_{16} \end{pmatrix} - \begin{pmatrix} m_0 \\ m_1 \\ m_2 \\ m_3 \end{pmatrix} \quad (20)$$

$$\begin{pmatrix} 1 & 0 & 0 & \cdots & & & & 0 \\ c_1 & 1 & 0 & & & & & \vdots \\ c_2 & c_1 & 1 & \ddots & & & & \\ c_3 & c_2 & c_1 & \ddots & \ddots & & & \\ c_4 & c_3 & c_2 & c_1 & 1 & \ddots & & \\ 0 & c_4 & c_3 & \cdots & & \ddots & 0 & \vdots \\ \vdots & \vdots & & & & & 1 & 0 \\ 0 & \cdots & 0 & c_4 & c_3 & c_2 & c_1 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ e'_1 \\ \vdots \\ \\ \\ \\ \vdots \\ e'_{15} \end{pmatrix}$$

$$- \begin{pmatrix} 1 & 0 & 0 & \cdots & & & & 0 \\ d_1 & 1 & 0 & & & & & \vdots \\ d_2 & d_1 & 1 & \ddots & & & & \\ d_3 & d_2 & d_1 & \ddots & \ddots & & & \\ d_4 & d_3 & d_2 & d_1 & 1 & \ddots & & \\ 0 & d_4 & d_3 & \cdots & & \ddots & 0 & \vdots \\ \vdots & \vdots & & & & & 1 & 0 \\ 0 & \cdots & 0 & d_4 & d_3 & d_2 & d_1 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ \widetilde{e}_1 \\ \vdots \\ \\ \\ \\ \vdots \\ \widetilde{e}_{15} \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ 0 \end{pmatrix} \quad (21)$$

With $\mathbf{x} = (d_1 \quad d_2 \quad d_3 \quad d_4 \quad \widetilde{e}_1 \quad \cdots \quad \widetilde{e}_{15})$, the following problem has been solved:

$F(\mathbf{x}) = 0$

With *F* defined by the matrix relations (20) and (21), and using a standard optimization routine.

# FAULT DETECTION BY MEANS OF DCS ALGORITHM COMBINED WITH FILTERS BANK

## *Application to the Tennessee Eastman Challenge Process*

Oussama Mustapha, Mohamad Khalil
*Lebanese University, Faculty of Engineering, Section I- El Arz Street, El Kobbe, Lebanon*
*Islamic University of Lebanon, Biomedical Department, Khaldé, Lebanon*
*oussama_mustapha@hotmail.com, mkhalil@ieee.org*

Ghaleb Hoblos, Houcine Chafouk
*ESIGELEC, IRSEEM, Saint Etienne de Rouvray, France*
*ghaleb.hoblos@esigelec.fr, houcine.chafouk@esigelec.fr*

Dimitri Lefebvre
*GREAH – University Le Havre, France*
*dimitri.lefebvre@univ-lehavre.fr*

Keywords: Signal processing, Filters Bank, Dynamic Cumulative Sum, Fault detection, Chemical processes.

Abstract: Early fault detection, which reduces the possibility of catastrophic damage, is possible by detecting the change of characteristic features of the signals. The aim of this article is to detect faults in complex industrial systems, like the Tennessee Eastman Challenge Process, through on-line monitoring. The faults that are concerned correspond to a change in frequency components of the signal. The proposed approach combines the filters bank technique, for extracting frequency and energy characteristic features, and the Dynamic Cumulative Sum method (DCS), which is a recursive calculation of the logarithm of the likelihood ratio between two local segments. The method is applied to detect the perturbations that disturb the Tennessee Eastman Challenge Process and may lead the process to shut down.

# 1 INTRODUCTION

The fault detection and isolation (FDI) methods are of particular importance in industry as long as the early fault detection in industrial systems reduces the personal damages and economical losses. Basically, model-based and data-based methods can be distinguished for diagnosis purposes. Model-based diagnosis requires a sufficiently accurate mathematical model of the process and compares the measured data with the knowledge, provided by the model of the considered system, in order to detect and isolate the faults that disturb the process. Parity space approach, observers design and parameters estimators are well known examples of model-based methods (Blanke and al., 2003; Patton and al., 2000). In contrast, non-model-based diagnosis requires a lot of process measurements and can also be divided into signal processing methods and artificial intelligence approaches. This study continues our research in frequency domain, concerning fault detection by means of filters bank (Mustapha and al., 2007; Mustapha and al., 2007b). The aim of this article is to propose a method for the on-line detection of changes applied after a filters bank decomposition that is needed to explore the frequency and energy components of the signal. The Moving Average (MA) and Auto Regressive Moving Average (ARMA) band pass filters are used to explore the frequency components. The motivation is that the filters bank modeling can transform the frequency changes into energy changes. Then, the Dynamic Cumulative Sum detection method (Khalil and Duchêne, 2000) is applied to the filtered signals (sub-signals) in order to detect any change in the signal. Filters bank is preferred in comparison with wavelet transform (Mustapha and al., 2007) because it could be directly implemented as a real time method.

## 2 PROBLEM STATEMENT

This work is originated from the analysis and characterization of random signals. In our case, the recorded signals can be described by a random process $x(t)$ as $x(t) = x_1(t)$ before the point of change $t_r$ and $x(t) = x_2(t)$ after the point of change $t_r$ where $t_r$ is the real time of detection. $x_1(t)$ and $x_2(t)$ can be considered as random processes where the statistical features are unknown but assumed to be identical for each segment 1 or 2. Therefore we assume that the signals $x_1(t)$ and $x_2(t)$ have Gaussian distributions. We will suppose also that the appearance times of the changes are unpredictable. We also suppose that the frequency distribution is an important factor for discriminating between the successive events.

Knowing that the signals from industrial systems are considered as slowly varying non-stationary ones, each change could be identified by its frequency content; our approach assumes piecewise stationary signals and the statistical parameters are the same for the two segments before and after the change. The application of any sequential detection algorithm directly on the original signal will decrease the probability of detection. However, after filters bank decomposition, the frequency change will be transformed into energy change and the detectability of the sequential detection algorithm will be improved. After decomposition of $x(t)$ into $N$ components : $y^{(m)}(t)$, $m = 1,..,N$, the problem of detection can be transformed to an hypothesis test: $H_0 : y^{(m)}(t)$, $t \in \{1,\ldots, t_r\}$ has a probability density function $f_0$ and $H_1 : y^{(m)}(t)$, $t \in \{ t_r + 1,\ldots, n\}$ has a probability density function $f_1$.

## 3 FILTERS BANK TECHNIQUE

In order to explore the frequency and energy components of the original signal, an important pre-processing step is required before detection, feature extraction and classification. At a discrete time $t$, the signal is first decomposed by using an N-channels band-pass filters bank whose central frequency moves from lowest frequency $f_1$ up to the highest frequency $f_N$. Each component $m \in \{1, \ldots, N\}$ is the result of filtering the original signal $x(t)$ by a band-pass filter centered on $f_m$. The frequency response curves of the filters bank is shown in figure 1. $f_N$ must satisfy the condition $f_N \leq f_s /2$, $f_s$ is the sampling frequency of the original signal $x(t)$, $N$ is the number of channels used. The choice of the filters bank depends on the original signal and its frequency band. The number of filters $N$ depends on the details that we have to extract from the signal and to the events that must be distinguished.

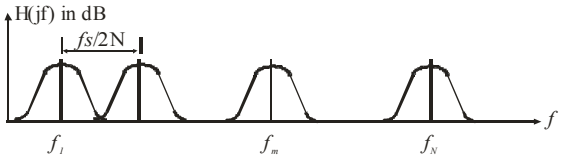



Figure 1: Responses of the filters bank.

The procedure of decomposing $x(t)$ into signals $y^{(m)}(t)$, $m = 1,..,N$, allows us to explore all frequency components of the signal. $y^{(1)}(t)$ gives the low frequency components and $y^{(N)}(t)$ gives the high frequency ones. Therefore, the points of change of each component give information about the frequency and energy contents and will be used to detect any changes in frequency and energy in the original signal.

For each component $m$, and at any discrete time $t$, the sample $y^{(m)}(t)$ of an ARMA-type filter is on-line computed according to the original signal $x(t)$ and using the parameters $a_i^{(m)}$ and $b_i^{(m)}$ of the corresponding band-pass filter according to the difference equation (1):

$$y^{(m)}(t) = \sum_{i=0}^{p} b(i)^{(m)} x(t-i) - \sum_{i=1}^{q} a(i)^{(m)} y^{(m)}(t-i) \qquad (1)$$

where $x(t)$ is the input signal of the filter, $y^{(m)}(t)$ is the output signal from the filter $m$, $a^{(m)}(i)$ and $b^{(m)}(i)$ are the numerator / denominator coefficients of the filter at level $m$, $a^{(m)}(0) = 1$, $m=1,...,N$, and $p$ and $q$ are the orders of the filter for a given level $m$, and they are assumed to be identical at any level, for simplicity.

The result of detection depends on the number of the band pass filters used, the central frequencies and the bandwidth of each channel. In practice, filters are uniformly chosen between zero Hertz and the half of the sampling frequency ($fs/2$). For real applications, the choice of the band pass filters are done after comparing the spectral density of two segments (signals $x_1(t)$ and $x_2(t)$). We start with $N$ filters and then reject the filters that do not give energy changes in sub-signals. The technique of comparing the frequency content (deciles or percentiles) is used by many authors to select the best filters (Falou, 2002).

# 4 SEQUENTIAL ALGORITHMS OF DETECTION

## 4.1 Cumulative Sum Method

The Cumulative Sum algorithm (CUSUM) is based on a recursive calculation of the logarithm of the likelihood ratios (Basseville and Nikiforov, 1993; Nikiforov, 1986). Let $x_1,x_2,x_3,\ldots,x_t$ be a sequence of observations. Let us assume that the distribution of the process $X$ depends on parameter $\theta_0$ until time $t_r$ and depends on parameter $\theta_1$ after the time $t_r$. At each time $t$ we compute the sum of logarithms of the likelihood ratios as follows:

$$S_1^{(t,m)} = \sum_{i=1}^{t} s^{(m)}{}_i = \sum_{i=1}^{t} Ln \frac{f_{\theta_1}(x_t / x_{t-1},\ldots,x_1)}{f_{\theta_0}(x_t / x_{t-1},\ldots,x_1)} \quad (2)$$

where, $f_\theta$ is the probability density function. The importance of this sum comes from the fact that its sign changes after the point of change. The real point of change $t_r$ can be estimated by $t_c$:

$$t_c = \max\{t : S_1^{(t,m)} - \min\{i : S_1^{(i,m)}\} = 0\}. \quad (3)$$

## 4.2 Dynamic Cumulative Sum Method

The Dynamic Cumulative Sum method (DCS) is based on the local dynamic cumulative sum, around the point of change $t_r$, and can be used when the parameters of the signal are unknown (Khalil and Duchêne, 2000). It is based on the local cumulative sum of the likelihood ratios between two local segments estimated at the current time $t$. These two dynamic segments $S_a^{(t)}$ (after $t$) and $S_b^{(t)}$ (before $t$) are estimated by using two windows of width $W$ (figure 2) before and after the instant $t$ as follows:

- $S_b^{(t)} : x_i ; \mathrm{i} = \{t-W,\ldots,t-1\}$ follows a probability density function $f_{\theta_b}(x_i)$
- $S_a^{(t)} : x_i ; \mathrm{i} = \{t+1,\ldots,t+W\}$ follows a probability density function $f_{\theta_a}(x_i)$

The parameters $\hat{\theta}_b^{(t)}$ of the segment $S_b^{(t)}$, are estimated using $W$ points before the instant $t$ and the parameters $\hat{\theta}_a^{(t)}$ of the segment $S_a^{(t)}$, are estimated using $W$ points after the instant $t$. At a time $t$, and for each level $m$, the DCS is defined as the sum of the logarithm of likelihood ratios from the beginning of the signal up to the time $t$:

$$DCS^{(m)}(S_a^{(t)}, S_b^{(t)}) = \sum_{i=1}^{t} Ln \frac{f_{\hat{\theta}_a}^{(i)}(x_i)}{f_{\hat{\theta}_b}^{(i)}(x_i)} = \sum_{i=1}^{t} \tilde{s}_i \quad (4)$$

(Khalil, 1999) proves that the DCS function reaches its maximum at the point of change $t_r$. The detection function used to estimate the point of change is:

$$g^{(m)}{}_t = \max_{1\le i\le t}\left[DCS^{(m)}\left(S_a^{(t)}, S_b^{(t)}\right)\right] - DCS^{(m)}\left(S_a^{(t)}, S_b^{(t)}\right) \quad (5)$$

The instant at which the procedure is stopped is $t_s = min\ \{t : g^{(m)}{}_t \ge h\}$, where $h$ is the detection threshold. The point of change is estimated as follows:

$$t_c = \max\{t>1 : g^{(m)}{}_t = 0\} \quad (6)$$

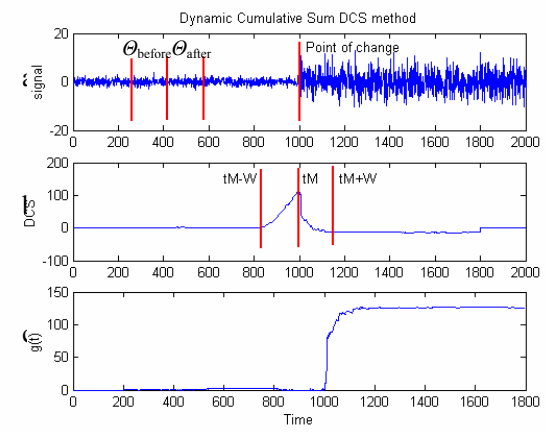



Figure 2: Application of the DCS on a signal of abrupt change. a) Original signal; b) DCS function; c) Detection function $g(t)$.

## 4.3 DCS Algorithm Combined with MA-type Filters Bank Decomposition

The detection is improved when the DCS method is applied after ARMA or MA modeling, especially when the signal presents no abrupt change, and the direct application of the DCS algorithm leads to ambiguous results that are sometimes difficult to interpret for accurate fault detection. In case of MA modeling, (i.e. $a_i = 0$), equation (2) leads to (7):

$$y^{(m)}(t) = \sum_{i=0}^{p} b(i)^{(m)} x(t-i) \quad (7)$$

In (Mustapha and al., 2007b), the detectability of the DCS algorithm after MA – type filters bank is proved. The basic idea is to prove that a change in a parameter is equivalent to a change of the sign of the expectation of the logarithm of the likelihood ratio: before the instant of change, $E(\tilde{S}_t) > 0$ and after instant of change $E(\tilde{S}_t) < 0$ where :

$$\tilde{S}_t = Ln(S_t) = \frac{1}{2}[Ln\frac{(\sigma_a^2)^{(t)}}{(\sigma_b^2)^{(t)}} + x_t^2(\frac{1}{(\sigma_a^2)^{(t)}} - \frac{1}{(\sigma_b^2)^{(t)}})] \quad (8)$$

and $(\sigma_a)^{(t)}$ stands for the variance of the segment $S_a^{(t)}$ and $(\sigma_b)^{(t)}$ for the variance of the segment $S_b^{(t)}$. For MA filter and assuming that the successive samples of *x(t)* are independent, (8) leads to (9):

$$E\left[\tilde{S}_t\right] = E\left[Ln(S_t)\right] = E\left[\frac{1}{2}Ln\frac{(\gamma_a^{(t)})^2}{(\gamma_b^{(t)})^2} + \frac{1}{2}\sum_{i=0}^{n} b^2(i)x^2(t-i)\frac{1}{(\gamma_a^{(t)})^2} - \frac{1}{2}\sum_{i=0}^{n} b^2(i)x^2(t-i)\frac{1}{(\gamma_b^{(t)})^2}\right] \quad (9)$$

For $t < t_r - W$, $S_a^{(t)}$ and $S_b^{(t)}$ are identical and have the same characteristics so, $E(\tilde{S}_t) = 0$. For $t_r - W < t < t_r$, $S_a^{(t)}$ and $S_b^{(t)}$ are no longer identical and $E(\tilde{S}_t) > 0$, and for $t_r < t < t_r + W$ we have $E(\tilde{S}_t) < 0$. Finally for $t > t_r + W$, $S_a^{(t)}$ and $S_b^{(t)}$ are identical again and $E(\tilde{S}_t) = 0$.

This demonstrates that $\tilde{S}_t$ increases before $t_r$, reaches a maximum at $t_r$ then decreases. So, in order to detect the point of change $t_r$, we search to detect the maximum of $\tilde{S}_t$ by using the detection function $g_t$.

# 5 FUSION TECHNIQUE

Because the detection algorithm is applied individually to each frequency component, it is important to apply a fusion technique to the resulting times of change in order to get a single value for a given fault in the system. The fusion technique is achieved as follows:

- -Each point of change at a given level is considered as an interval $[t_c-a, t_c+a]$, where $a$ is an arbitrary number of points taken before and after the point of change.
- -All the time intervals that have a common time area are considered to correspond to the same fault.
- -The resulting point of change $t_f$ is calculated as the center of gravity (or mean) of the superimposed intervals.

# 6 APPLICATION TO TECP

In this section, the method, based on filters bank decomposition and DCS algorithm, is applied to detect disturbances on the Tennessee Eastman Challenge Process (TECP; Downs and Vogel, 1993). The TECP is a multivariable non-linear, high dimensionality, unstable open-loop chemical reactor, that is a simulation of a real chemical plant provided by the Eastman company. There are 20 disturbances IDV1 through IDV20 that could be simulated (Downs and Vogel, 1993; Singhal, 2000). The sampling period for measurements is 60 seconds.

The TECP offers numerous opportunities for control and fault detection and isolation studies. In this work, we use a robust adaptive multivariable (4 inputs and 4 outputs) RTRL neural networks controller (Leclercq and al., 2005; Zerkaoui and al., 2007) This controller compensates all perturbations IDV1 to IDV 20 excepted IDV1, IDV6 and IDV7.

The figure 3 illustrates the advantage of our method to detect changes for real world FDI applications. Measurements of the reactor temperature (figure 3a) are decomposed into 3 components and according a 3 – channels band pass filters bank (figure 3c, d, e). The sampling frequency of this signal is 0.0167 Hz and the normalized central frequencies of the filters are: $f_{c1}$ = 0.64, $f_{c2}$ = 0.74, $f_{c3}$ = 0.77. From time $t_r$ = 600 hours, the unknown perturbation IDV16 modifies the dynamic behavior of the system. The detection functions applied on the 3 components (figure 3f, g, h) can be compared with the detection function applied directly on the measurement of pressure (figure 3b).

The detection results are considerably improved by using the filters bank as a -preprocessing. In that case, the DCS applied on original signal is not suitable to detect the perturbation whereas the DCS combined with 3- channels band pass filters bank can detect the perturbation. After fusion, the estimated instant of change is $t_f$ = 669 hours that include a large delay to detection of 69 hours.

Table 1: Detection delays for several perturbations in TECP.

| Disturbance | Significance | *T°* | *Pr* | *sepL* | *StrL* |
|---|---|---|---|---|---|
| IDV 2 | B composition, A/C ratio constant (step) | 599/677 | 601/ 665 | 510/ 535 | 502/518 |
| IDV 3 | D feed temperature (step) | 665/680 | X | X | X |
| IDV 4 | Reactor cooling water inlet temperature (step) | 602/603 | X | X | X |
| IDV 8 | A, B, C feed composition (random variation) | 650/660 | 650/660 | 513/634 | 343/353 |
| IDV 9 | D feed temperature (random variation) | 279/287 | X | X | X |
| IDV 11 | Reactor cooling water inlet temperature (random variation) | 607/608 | X | X | X |
| IDV 16 | Unknown | 647/670 | X | X | X |
| IDV 17 | Unknown | 660/850 | X | X | X |

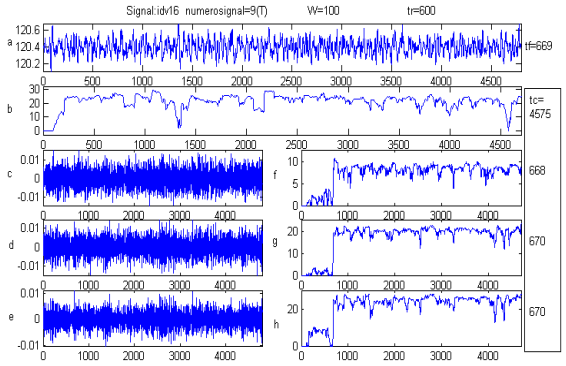



Figure 3: Analysis of the reactor temperature measurements (°C) for TECP with robust adaptive control and for IDV 16 perturbation from $t_r$ = 600. a) Original signal b) DCS applied directly on the original signal c) d) e)Decomposition using band pass filters($m$ = 1,2,3) f) g) h) Detection functions applied on the filtered signals (c, d, e).

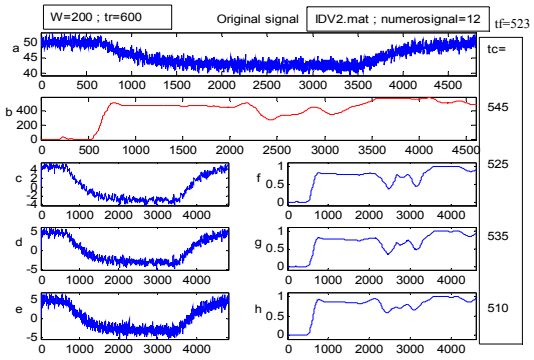



Figure 4: Analysis of the reactor separator level (%) for TECP with robust adaptive control and for IDV 2 perturbation from $t_r$ = 600. a) Original signal b) DCS applied directly on the original signal c) d) e)Decomposition using band pass filters ($m$ = 1,2,3) f) g) h) Detection functions applied on the filtered signals (c, d, e).

The diagnosis of numerous perturbations has been investigated with our method in order to show the efficiency of the approach. All perturbations have been simulated starting from time $t_r$ = 600 hours. The table 1 shows the results obtained with various measured signals and various perturbations. Two studies have been considered:

- For perturbations IDV 2 – 3 – 4 – 8 – 9 – 11 –16 –17, the detection has been investigated in a systematic way from the measurements of temperature in reactor.
- For perturbations IDV 2 and IDV 9, the detection has been compared depending on the measured variable (*T, Pr, StrL, SepL*).

Table 1 shows the minimal and maximal values of $t_c$ obtained over the three components. The detection of changes was satisfactory in most cases depending on the measured signals and the filters that have been used. It is already important to notice that IDV 2, that consists in a step in B composition, cannot be detected with $Y_3$ and $Y_4$ (dark grey cells). This perturbation corresponds to a modification of the mean value that can be detected with other methods (figure 4). IDV 8 and IDV 9 also present some difficulties with some measured variable. But an adaptation of threshold $h$ used with detection function and an adaptation of the central decomposition frequencies will lead to acceptable results. One can also notice the large dispersion of the detection times in some cases.

# 7 CONCLUSIONS

The aim of our work is to detect the point of change of statistical parameters in signals issued from complex industrial processes. This method uses a band-pass filters bank combined with DCS to characterize and classify the parameters of a signal in order to detect any variation of the statistical parameters due to any change in frequency and energy. The proposed algorithm provides good results for the detection of frequency changes in the signal and can be used to detect the perturbation of chemical processes as the TECP under stable closed loop control. The results illustrate the interest of the approach for on – line detection and real world applications. Changes due to faults are easily separated from changes due to input variations by the comparative analysis of input and output signals.

In the future, we will investigate detectability in case of abrupt variations of the mean (figure 4). We will also consider multiple faults investigation and fault isolation based on signatures table of faults. Fault isolation can be studied according to the classification of the changes that are detected and can certainly be improved by increasing the number of considered filters and adapting their central frequencies. We will also study the automatic adaptation of the detection threshold *h* and complete the diagnosis with faults identification.

# REFERENCES

Basseville M., Nikiforov I. *Detection of Abrupt Changes: Theory and Application*. Prentice-Hall, Englewood Cliffs, NJ, 1993.

Blanke M., Kinnaert M., Lunze J., Staroswiecki M., *Diagnosis and fault tolerant control*, Springer Verlag, New York, 2003.

Downs, J.J., Vogel, E.F, A plant-wide industrial control problem, *Computers and Chemical Engineering,* 17, pp. 245-255, 1993.

Falou W., "Une approche de la segmentation dans des signaux de longue durée fortement bruités. Application en ergonomie", PhD. Thesis, Université de Technologie de Troyes, France, 2002.

Khalil M., Une approche pour la détection fondée sur une somme cumulée dynamique associée à une décomposition multiéchelle. Application à l'EMG utérin. *Dix-septième Colloque GRETSI sur le traitement du signal et des images*, Vannes, France,1999.

M. Khalil, J. Duchêne. Uterine EMG Analyzing: A dynamic approach for change detection and classification. *IEEE Transactions on Biomedical Engineering, vol. 46, N6, pp. 748-756, juin 2000.*

Leclercq, E., Druaux, F. Lefebvre, D., Zerkaoui, S., Autonomous learning algorithm for fully connected recurrent networks. *Neurocomputing*, vol. 63, pp. 25-44, 2005.

Mustapha O., Khalil M., Hoblos G, Chafouk H., Ziadeh H., Lefebvre D., About the Detectability of DCS Algorithm Combined with Filters Bank*, Qualita 2007*, Tanger, Maroc, 2007.

Nikiforov I. Sequential detection of changes in stochastic systems. *Lecture notes in Control and information Sciences*, NY, USA, pp. 216-228, 1986.

Patton R.J., Frank P.M. and Clarck R., *Issue of Fault diagnosis for dynamic systems*, Springer Verlag, 2000.

Singhal, A., Tennessee Eastman Plant Simulation with Base Control System of McAvoy and Ye., *Research report, Department of Chemical Engineering,* University of California, Santa Barbara, USA, 2000.

Zerkaoui S., Druaux F., Leclercq E., Lefebvre D., Multivariable adaptive control for non-linear systems : application to the Tennessee Eastman Challenge Process, *ECC 2007*, Kos, Greece, 2007.

# DIRECTIONAL CHANGE IN A PRIORI ANTI-WINDUP COMPENSATORS VS. PREDICTION HORIZON

Dariusz Horla
*Poznan University of Technology, Institute of Control and Information Engineering*
*Division of Control and Robotics, ul. Piotrowo 3a, 60-965, Poland*
*Dariusz.Horla@put.poznan.pl*

Keywords: Directional change, windup phenomenon, optimal control, linear matrix inequalities, predictive control.

Abstract: The paper presents the correspondence in between directional change and anti-windup phenomenon with respect to a priori anti-windup compensator on the basis of MPC (simulation results include plants with not equal number of inputs and outputs). It shows what is the excess of directional change for consecutive predictions of control vectors for a given prediction horizons.

## 1 INTRODUCTION

Taking control limits into consideration is necessary to achieve high performance of the designed control systems (Horla, 2006b). There are two ways in which one can consider possible constraints at synthesis of controllers. In the first approach, imposing constraints during the design procedure of the controller usually leads to difficulties with obtaining explicit form of control laws, apart from very simple cases. The other way is to assume the system is fully linear and, subsequently, having designed the controller for unconstrained system (by means of optimisation, using Diophantine equations, etc) – impose constraints, what would require additional changes in control system due to presence of constraints (Horla, 2007b; Öhr, 2003; Peng et al., 1998).

The situation when because of constraints internal controller states do not correspond to the actual signals present in the control systems is referred in the literature as windup phenomenon (Öhr, 2003). One can expect inferior performance because of infeasibility of computed (unconstrained) control signals when control limits are not taken into account.

A few methods of compensating the windup phenomenon from SISO framework work well enough in the case of multivariable systems (Öhr, 2003; Walgama and Sternby, 1993). In such a case, apart from the windup phenomenon itself, one can also observe directional change in the control vector due to different implementations of constraints, what could affect direction of the unconstrained control vector (Horla, 2004; Horla, 2007a).

The other problem is, in general form, decoupling, with respect to not equal number of control signals and output signals, when control direction corresponds not only to input principal directions or maximal directional gain of the transfer function matrix, but also to the degree of decoupling (Albertos and Sala, 2004; Maciejowski, 1989).

The problem of directional change has been initially discussed in (Walgama and Sternby, 1993). The paper (Horla, 2007a) defined the connection of directional change problem with anti-windup compensation (AWC) for systems with equal numer of inputs and outputs.

The current paper has been given rise by research carried out in (Horla, 2004; Horla, 2007a; Horla, 2007b) and extends the understanding of anti-windup compensation to non-square systems with imposed constraints, comparing control performance with optimisation-based approach, related to MPC (Camacho and Bordons, 1999; Doná et al., 2000; Maciejowski, 2002) that is widely-spread and applied in the industry. In the paper, the problem of directional change has been studied with respect to optimal a priori anti-windup compensation and different prediction horizons.

## 2 A PRIORI AWC

One can perform anti-windup compensation by incorporating AWC implicitly into the controller. In order to use all the advantages of such an approach (as optimality of the solution, no need to design decoupling stages, etc.), let the optimal constrained control vector

follow from

$$\underline{\underline{u}}_t^{\star} : J_t(\underline{\underline{u}}_t^{\star}) = \inf_{\underline{\underline{u}}_t \in \mathscr{D}(J_t)} \left\{ J_t(\underline{\underline{u}}_t) \right\}, \tag{1}$$

with constraints

$$|u_{j,i,t}| \quad \leq \quad \alpha_j \,, \tag{2}$$

where $\alpha_j$ is an amplitude constraint of the $i$–th element of control vector, and $\underline{u}_{j,t}$, $\mathscr{D}(J_t)$ is the set of all control vectors such that $J_t$ has a finite value,

$$\underline{\underline{u}}_t = \begin{bmatrix} \underline{u}_{1,t} \\ \underline{u}_{2,t} \\ \vdots \\ \underline{u}_{m,t} \end{bmatrix}, \tag{3}$$

and $\underline{u}_{j,t}$ $(1 \leq j \leq m)$ comprises sequences of control actions applied to the $j$–th input with control horizon $N_u$ (Horla, 2006a).

The controller is responsible for tracking given reference (implicit model) output vector $\underline{r}_{M,t}$ with $\underline{y}_t$ minimising

$$J_t \quad = \quad \sum_{k=1}^{p} \sum_{l=d}^{d+N_u-1} (r_{M,k,t+l} - y_{k,t+l})^2, \tag{4}$$

which can be presented in the sense of $L_2$ norm as

$$J_t \quad = \quad \left\| \underline{\underline{r}}_{M,t+d} - \underline{\hat{y}}_{t+d} - \underline{\hat{\hat{y}}}_{t+d} \right\|_2^2, \tag{5}$$

where

$$\underline{\underline{r}}_{M,t+d} = \begin{bmatrix} \underline{r}_{M,1,t+d} \\ \underline{r}_{M,2,t+d} \\ \vdots \\ \underline{r}_{M,p,t+d} \end{bmatrix} \tag{6}$$

comprises vectors including reference signals known for $d+N_u-1$ steps in advance.

The vector of prediction of plant response forced by the sought control sequence is computed in an iterative manner

$$\underline{\hat{y}}_{t+d} \quad = \quad \begin{bmatrix} \underline{\hat{y}}_{1,t+d} \\ \underline{\hat{y}}_{2,t+d} \\ \vdots \\ \underline{\hat{y}}_{p,t+d} \end{bmatrix}, \tag{7}$$

where $\underline{\hat{y}}_{i,t+d} = G\underline{\underline{u}}_t$, and $G$ comprises matrices of plant impulse response samples.

The controller is to search for control vectors $\underline{u}_{j,t}$ with $1 \leq j \leq m$, each of them being control sequences applied to the $j$–th input $\{u_{j,t}, u_{j,t+1}, \ldots, u_{j,t+N_u-1}\}$ in horizon $N_u > 0$. Based on the superposition rule, the decay-response vector $\underline{\hat{\hat{y}}}_{t+d}$ subject to initial conditions $\underline{u}_{t-k}$ $(k \geq 0)$ is computed iteratively alike.

Having expressed (4) as

$$J_t = \left( G\underline{\underline{u}}_t + \underline{\hat{y}}_{t+d} - \underline{\underline{r}}_{M,t+d} \right)^T \left( G\underline{\underline{u}}_t + \underline{\hat{y}}_{t+d} - \underline{\underline{r}}_{M,t+d} \right), \tag{8}$$

minimisation subject to constraints is equivalent to (Boyd et al., 1994; Boyd and Vandenberghe, 2004)

$$\begin{aligned} &\min \gamma \\ &\text{s.t.} \\ &\begin{bmatrix} I & \star \\ \underline{\underline{u}}_t^T (G^T G)^{1/2} & \begin{pmatrix} \gamma - (\underline{\underline{r}}_{M,t+d} - \underline{\hat{\hat{y}}}_{t+d})^T \times \\ \times (\underline{\underline{r}}_{M,t+d} - \underline{\hat{\hat{y}}}_{t+d}) + \\ + 2(\underline{\underline{r}}_{M,t+d} - \underline{\hat{\hat{y}}}_{t+d})^T G\underline{\underline{u}}_t \end{pmatrix} \end{bmatrix} \geq 0, \\ &\operatorname{diag}\left\{ F^{(1,1)}, \ldots, F^{(1,m)} \right\} \geq 0 \geq 0, \\ &\operatorname{diag}\left\{ F^{(2,1)}, \ldots, F^{(2,m)} \right\} \geq 0, \end{aligned} \tag{9}$$

where the last to LMIs define upper and lower bounds of $\underline{\underline{u}}_t$, and $\star$ is a symmetrical entry.

## 3 SIMULATION STUDIES

The following multivariable CARMA plant model will be of interest

$$A(q^{-1})\underline{y}_t = B(q^{-1})\underline{u}_{t-d} \,, \tag{10}$$

with left co-prime polynomial matrices $A(q^{-1})$, $B(q^{-1})$, delay $d = 1$, with $\underline{y}_t \in \mathscr{R}^p$ as the output vector, $\underline{u}_t \in \mathscr{R}^m$ is the constrained control vector ($\underline{v}_t \in \mathscr{R}^m$ will denote unconstrained control vector). The considered plants are assumed to be cross-coupled:

- P1 ($m = 2$, $p = 2$)

$$\begin{aligned} A(q^{-1}) \quad &= \quad I + \begin{bmatrix} 0.8 & -0.1 \\ 0.4 & -1.0 \end{bmatrix} q^{-1} + \\ &\quad + \begin{bmatrix} -0.49 & -0.10 \\ 0.10 & 0.25 \end{bmatrix} q^{-2}, \\ B(q^{-1}) \quad &= \quad \begin{bmatrix} 1.0 & 0.3 \\ 0.5 & 0.8 \end{bmatrix}, \end{aligned}$$

- P2 ($m = 3$, $p = 2$)

$$\begin{aligned} A(q^{-1}) \quad &= \quad I + \begin{bmatrix} 0.8 & -0.1 \\ 0.4 & -1.0 \end{bmatrix} q^{-1} + \\ &\quad + \begin{bmatrix} -0.49 & -0.10 \\ 0.10 & 0.25 \end{bmatrix} q^{-2}, \\ B(q^{-1}) \quad &= \quad \begin{bmatrix} 1.0 & 0.2 & 0.3 \\ 0.5 & 0.3 & 0.8 \end{bmatrix}, \end{aligned}$$

- P3 ($m = 2$, $p = 3$)

$$A(q^{-1}) = I + \begin{bmatrix} -0.7 & 0.0 & 0.1 \\ -0.1 & -0.8 & 0.2 \\ 0.1 & 0.0 & -0.8 \end{bmatrix} q^{-1} + \\ + \begin{bmatrix} -0.1 & 0.0 & 0.0 \\ 0.0 & 0.1 & 0.0 \\ 0.0 & 0.0 & 0.5 \end{bmatrix} q^{-2},$$

$$B(q^{-1}) = \begin{bmatrix} 1.0 & 0.1 \\ 0.2 & 1.0 \\ 0.5 & -0.1 \end{bmatrix}.$$

The reference vector is pre-filtered by implicit reference model with characteristic polynomial matrix

$$A_M(q^{-1}) = (1 - 0.5q^{-1})I^{p\times p},$$

what corresponds to closed-loop tracking with dynamics described by $A_M(q^{-1})$.

Evaluation of control performance connected with anti-windup compensation quality requires following performance indices to be introduced:

$$J_1 = \frac{1}{N}\sum_{i=1}^{p}\sum_{t=1}^{N}|r_{i,t} - y_{i,t}|, \quad (11)$$

$$J_2 = \frac{1}{N}\sum_{i=1}^{p}\sum_{t=1}^{N}(r_{i,t} - y_{i,t})^2, \quad (12)$$

$$\overline{\varphi}_i = \frac{1}{N}\sum_{t=1}^{N}|\varphi(\underline{y}_{t,i}) - \varphi(\underline{u}_{t,i})| \ [^\circ], \quad (13)$$

$$\overline{\varphi}_i^2 = \frac{1}{N}\sum_{t=1}^{N}(\varphi(\underline{y}_{t,i}) - \varphi(\underline{u}_{t,i}))^2, \quad (14)$$

where (11) corresponds to mean absolute tracking error of $p$ outputs, (13) is a mean absolute direction change in between computed and constrained control vector, and $\varphi(i)$ denotes angle measure of control vector sequence in prediction horizon $N_u = i$.

## 4 SIMULATION RESULTS

For plants P1 and P2 the reference vectors comprise piecewise constant reference signals, whereas for P3 the third output is to be kept at zero at all times, what is difficult when there is a inferior number of control inputs in comparison with plant outputs.

Numerical results of performed simulations have been presented in Tables 1 and 2. The first set of simulations tested to what excess the directional change phenomenon will take place for plants P1–P3 and different prediction horizon.

As it can be seen from Table 1a and Figures 1 and 4, for P1, the greatest directional change (with respect to unconstrained control vector generated at the same time instant, but not applied) takes place in the current sample. The greater the prediction horizon, the smaller the directional change becomes. Since a mean angle deviations is approx. $1^\circ$ then, one can say that constrained control vector is close to the computed unconstrained control vector. This might also take place because of equal number of inputs and outputs, what leads to easier decoupling.

In the case of P2 (Tab. 1b, Fig. 2, 5), mean angle deviation is near the right angle, what corresponds to to normal vectors with the third component unchanged, i.e. rotation with respect to a fixed axis. This might be be connected with plant principal directions and with the need to decouple outputs from inputs. Since the number of control inputs is greater than plant outputs, the excessive change in direction is needed, because one can obtain better tracking performance than for $m = p = 2$.

If the plant has insufficient number of control inputs (P3, Tab. 2b, Fig. 3, 6), it is impossible to assure high control performance and one has to cope with potential problem of uncontrollable modes. As it can be seen, the speed of transients has been reduced, what lead to better decoupling, aiding anti-windup compensation. In such a case, often directional change is a result of the need of decoupling.

For the case of no directional change requirement (Tab. 2), such a regime of work (present in some applications in robotics, or e.g. in tracking, (Öhr, 2003)), results in inferior control performance. For P1 and increasing $N_u$ one obtains performance degradation, for P2 the closed-loop system becomes unstable (in order to decouple, the controller would have to alter control direction) the only improvement can be observed in the case of P3 because of $m < p$ (where some coupling is always present and results in proportions between control vector components that controller has to abide to).

## 5 SUMMARY

As it has been shown in the paper, the problem of directional change can be presented in a different way for plants with $m \neq p$ than in (Horla, 2007a; Walgama and Sternby, 1993). Not allowing directional change, may cause instability in the case of unstable plants (see P2), whereas for the other cases it degrades control performance.

Altering control direction is related to decoupling, thus one can expects problems with performance for $m > p$ and good control quality for $m < p$ when components of control vector must be kept in proportion (e.g., in a circular shape cutting task) at all times.

Table 1: a) $p=2$, $m=2$, b) $p=2$, $m=3$, c) $p=3$, $m=2$.

| $a)$ | $N_u=1$ | $N_u=2$ | $N_u=3$ | $N_u=4$ | $N_u=5$ |
|---|---|---|---|---|---|
| $J_1$ | 0.6354 | 0.6144 | 0.6034 | 0.5951 | 0.5911 |
| $J_2$ | 1.9002 | 1.7126 | 1.6219 | 1.5683 | 1.5338 |
| $\overline{\varphi}_1$ | 0.7171 | 0.8604 | 1.4556 | 1.5468 | 1.6631 |
| $\overline{\varphi}_2$ | | 0.7736 | 0.9795 | 1.1555 | 1.3119 |
| $\overline{\varphi}_3$ | | | 0.7398 | 1.0107 | 1.1649 |
| $\overline{\varphi}_4$ | | | | 0.7299 | 0.9942 |
| $\overline{\varphi}_5$ | | | | | 0.7207 |
| $\overline{\varphi}_1^2$ | 11.8327 | 9.5201 | 111.7057 | 115.1060 | 124.3589 |
| $\overline{\varphi}_2^2$ | | 10.2881 | 13.1625 | 19.5663 | 25.4658 |
| $\overline{\varphi}_3^2$ | | | 10.6466 | 15.1063 | 20.8268 |
| $\overline{\varphi}_4^2$ | | | | 10.6025 | 14.9204 |
| $\overline{\varphi}_5^2$ | | | | | 10.5455 |

| $b)$ | $N_u=1$ | $N_u=2$ | $N_u=3$ | $N_u=4$ | $N_u=5$ |
|---|---|---|---|---|---|
| $J_1$ | 0.3535 | 0.3518 | 0.3560 | 0.3581 | 0.3585 |
| $J_2$ | 0.7373 | 0.6985 | 0.6914 | 0.6923 | 0.6931 |
| $\overline{\varphi}_1$ | 95.1171 | 89.5872 | 89.5007 | 96.3644 | 93.5828 |
| $\overline{\varphi}_2$ | | 88.0821 | 90.1826 | 88.2265 | 86.9010 |
| $\overline{\varphi}_3$ | | | 88.9365 | 82.4423 | 91.6918 |
| $\overline{\varphi}_4$ | | | | 79.6011 | 87.4887 |
| $\overline{\varphi}_5$ | | | | | 102.5099 |
| $\overline{\varphi}_1^2$ | 9138.6 | 8218.2 | 8306.7 | 10250.0 | 9549.4 |
| $\overline{\varphi}_2^2$ | | 8341.1 | 8477.1 | 8938.1 | 8517.1 |
| $\overline{\varphi}_3^2$ | | | 8429.2 | 7331.5 | 9513.0 |
| $\overline{\varphi}_4^2$ | | | | 7492.2 | 8101.1 |
| $\overline{\varphi}_5^2$ | | | | | 11619.6 |

| $c)$ | $N_u=1$ | $N_u=2$ | $N_u=3$ | $N_u=4$ | $N_u=5$ |
|---|---|---|---|---|---|
| $J_1$ | 1.3293 | 1.2364 | 1.1721 | 1.1422 | 1.1407 |
| $J_2$ | 1.8375 | 1.3703 | 1.2450 | 1.1629 | 1.1177 |
| $\overline{\varphi}_1$ | 2.5244 | 3.6052 | 3.7189 | 3.8635 | 5.8012 |
| $\overline{\varphi}_2$ | | 1.8646 | 3.0450 | 3.2257 | 3.2991 |
| $\overline{\varphi}_3$ | | | 2.1306 | 3.4264 | 3.6449 |
| $\overline{\varphi}_4$ | | | | 2.0898 | 3.4420 |
| $\overline{\varphi}_5$ | | | | | 2.2384 |
| $\overline{\varphi}_1^2$ | 37.0812 | 109.3184 | 95.4019 | 106.2623 | 687.9498 |
| $\overline{\varphi}_2^2$ | | 33.4335 | 116.5809 | 121.1664 | 106.5284 |
| $\overline{\varphi}_3^2$ | | | 49.0232 | 135.6229 | 141.8349 |
| $\overline{\varphi}_4^2$ | | | | 46.1047 | 131.0049 |
| $\overline{\varphi}_5^2$ | | | | | 47.5444 |

Table 2: no directional change, a) $p=2$, $m=2$, b) $p=2$, $m=3$, c) $p=3$, $m=2$ ($-$ denotes unstable closed-loop system).

| $a)$ | $N_u=1$ | $N_u=2$ | $N_u=3$ | $N_u=4$ | $N_u=5$ |
|---|---|---|---|---|---|
| $J_1$ | 0.8846 | 0.8994 | 0.8674 | 0.8914 | 0.8975 |
| $J_2$ | 2.4978 | 2.6290 | 2.3793 | 2.3365 | 2.2729 |

| $b)$ | $N_u=1$ | $N_u=2$ | $N_u=3$ | $N_u=4$ | $N_u=5$ |
|---|---|---|---|---|---|
| $J_1$ | 9.5249 | 11.1171 | 10.6376 | – | – |
| $J_2$ | 36.8874 | 77.8512 | 62.5581 | – | – |

| $c)$ | $N_u=1$ | $N_u=2$ | $N_u=3$ | $N_u=4$ | $N_u=5$ |
|---|---|---|---|---|---|
| $J_1$ | 1.4536 | 1.4110 | 1.3620 | 1.3450 | 1.3510 |
| $J_2$ | 1.9418 | 1.6632 | 1.5240 | 1.4352 | 1.4242 |

# REFERENCES


Albertos, P. and Sala, A. (2004). *Multivariable Control Systems*. Springer-Verlag, London, United Kingdom.

Boyd, S., Ghaoui, L. E., Feron, E., and Balakrishnan, V. (1994). *Linear Matrix Inequalities in System and Control Theory*. Society for Industrial and Applied Mathematics, Philadelphia, United States of America, 3rd edition.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press, United Kingdom.

Camacho, E. and Bordons, C. (1999). *Model Predictive Control*. Springer-Verlag, United Kingdom.

Doná, J. D., Goodwin, G., and Seron, M. (2000). Anti-windup and model predictive control: Reflections and connections. *European Journal of Control*, 6(5):455–465.

Horla, D. (2004). Directional change and anti-windup compensation for multivariable systems. *Studies in Automation and Information Technology*, 28/29:53–68.

Horla, D. (2006a). LMI-based multivariable adaptive predictive controller with anti-windup compensator. In *Proceedings of the 12th IEEE International Conference MMAR*, pages 459–462, Miedzyzdroje.

Horla, D. (2006b). Standard vs. LMI approach to a convex optimisation problem in multivariable predictive control task with a priori anti-windup compensator. In *Proceedings of the 18th ICSS*, pages 147–152, Coventry.

Horla, D. (2007a). Directional change and windup phenomenon. In *Proceedings of the 4th IFAC International Conference on Informatics in Control Automation and Robotics*, pages CD–ROM, Angers, France.

Horla, D. (2007b). Optimised conditioning technique for a priori anti-windup compensation. In *Proceedings of the 16th International Conference on Systems Science*, pages 132–139, Wrocław, Poland.

Maciejowski, J. (1989). *Multivariable Feedback Design*. Addison-Wesley Publishing Company, Cambridge, United Kingdom.

Maciejowski, J. (2002). *Predictive Control with Constraints*. Pearson Education Limited, United Kingdom.

Öhr, J. (2003). *Anti-windup and Control of Systems with Multiple Input Saturations: Tools, Solutions and Case Studies*. PhD thesis, Uppsala University, Uppsala, Sweden.

Peng, Y., Vrančić, D., Hanus, R., and Weller, S. (1998). Anti-windup designs for multivariable controllers. *Automatica*, 34(12):1559–1565.

Walgama, K. and Sternby, J. (1993). Contidioning technique for multiinput multioutput processes with input saturation. *IEE Proceedings-D*, 140(4):231–241.

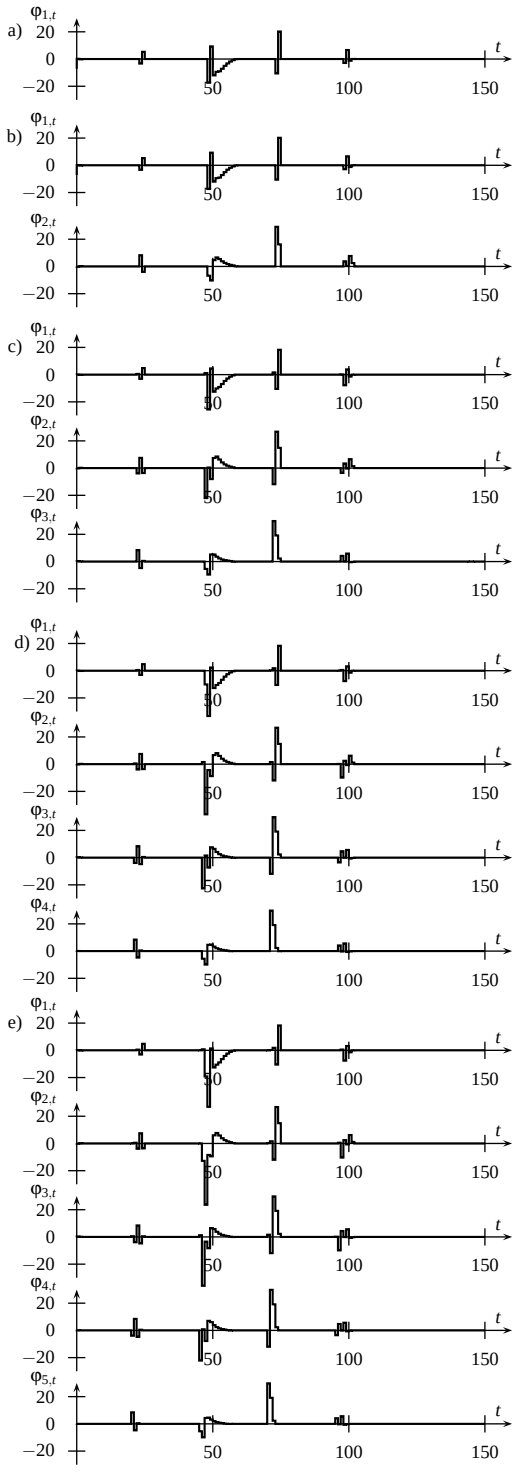



Figure 1: $p = 2$, $m = 2$, a) $N_u = 1$, b) $N_u = 2$, c) $N_u = 3$, d) $N_u = 4$, e) $N_u = 5$.

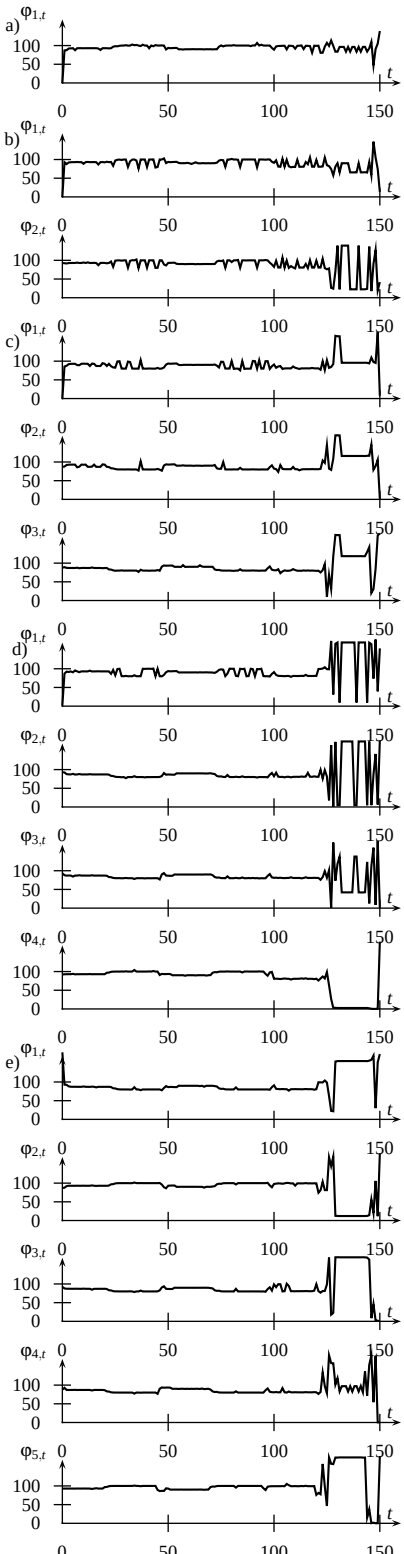



Figure 2: $p = 2$, $m = 3$, a) $N_u = 1$, b) $N_u = 2$, c) $N_u = 3$, d) $N_u = 4$, e) $N_u = 5$.

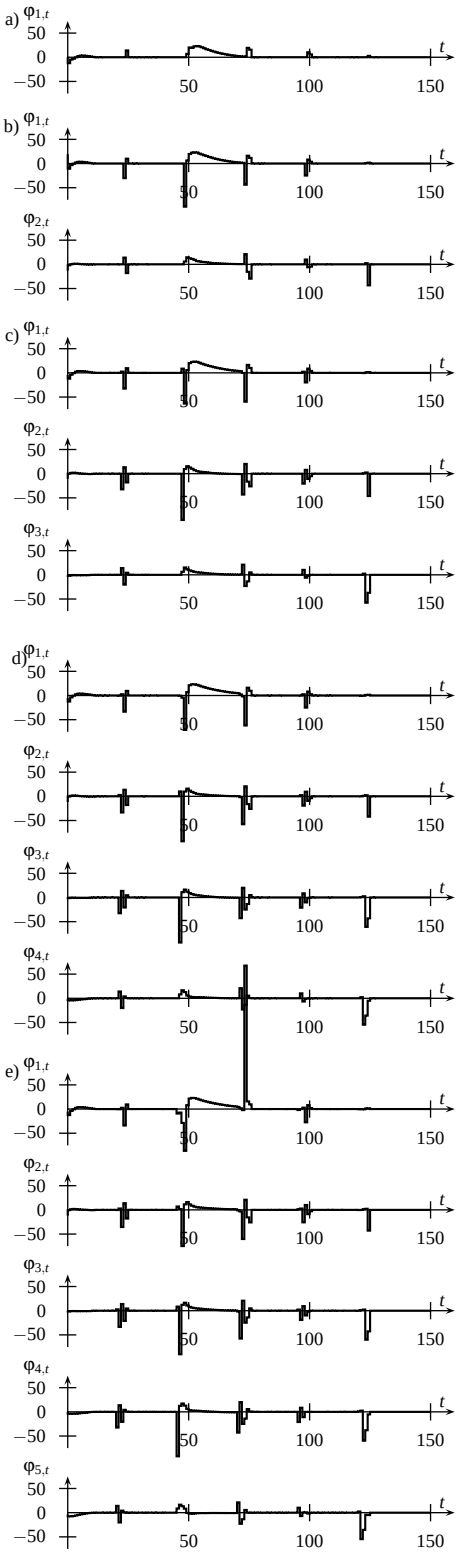



Figure 3: $p = 3$, $m = 2$, a) $N_u = 1$, b) $N_u = 2$, c) $N_u = 3$, d) $N_u = 4$, e) $N_u = 5$.

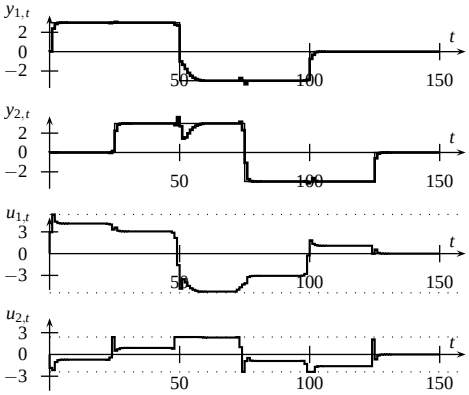



Figure 4: $p = 2$, $m = 2$, $N_u = 3$.

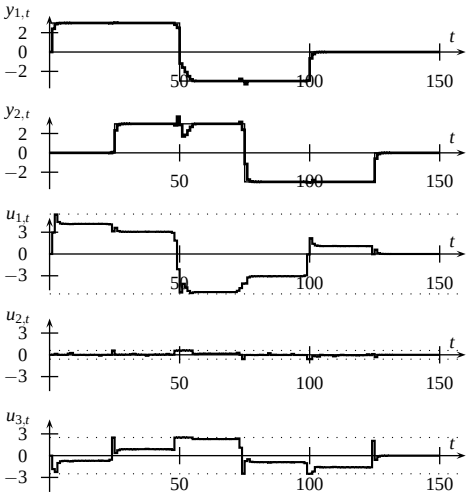



Figure 5: $p = 2$, $m = 3$.

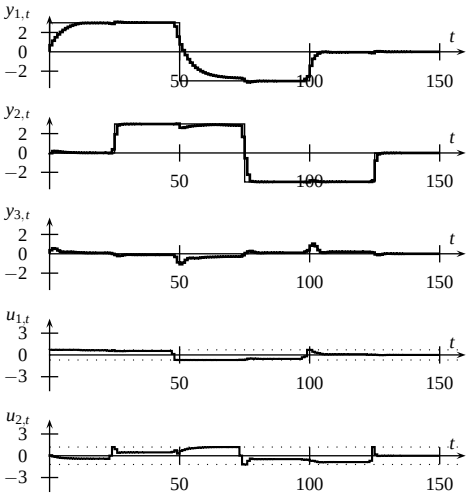



Figure 6: $p = 3$, $m = 2$.

# PHASE LOCKED LOOPS DESIGN AND ANALYSIS

Nikolay V. Kuznetsov, Gennady A. Leonov and Svetlana S. Seledzhi
*Saint-Petersburg State University, Universitetski pr. 28, Saint-Petersburg, 198504, Russia*
*leonov@math.spbu.ru*

Keywords: Mathematical model, stability, Phase-locked loops, Costas loop.

Abstract: New methods, for the design of different block diagrams of PLL, using the asymphtotic analysis of high-frequency periodic oscillations, are suggested. The PLL description on three levels is made: 1) on the level of electronic realizations; 2) on the level of phase and frequency relations between inputs and outputs in block diagrams; 3) on the level of differential and integro-differential equations. On the base of such description, the block diagram of floating PLL for the elimination of clock skew and that of frequency synthesizer is proposed. The rigorous mathematical formulation of the Costas loop for the clock oscillators are first obtained. The theorem on a PLL global stability is proved.

## 1 INTRODUCTION

The phase-locked loops are widespread in a modern radio electronics and circuit technology (Viterbi, 1966; Gardner, 1966; Lindsey, 1972; Lindsey and Chie, 1981; Leonov, Reitmann and Smirnova, 1992; Leonov, Ponomarenko and Smirnova, 1996; Leonov and Smirnova, 2000; Kroupa, 2003; Best, 2003, Razavi, 2003; Egan, 2000; Abramovitch, 2002). In this paper the technique of PLL description on three levels is suggested:

1) on the level of electronic realizations,

2) on the level of phase and frequency relations between inputs and outputs in block-diagrams,

3) on the level of differential and integro-differential equations.

The second level, involving the asymptotical analysis of high-frequency oscillations, is necessary for the well-formed derivation of equations and for the passage on the third level of description. For example, the main for the PLL theory notion of phase detector is formed exactly on the second level of consideration. In this case *the characteristic of phase detector depends on the class of considered oscillations*. While in the classical PLL it is used the oscillation multipliers, for harmonic oscillations, the characteristic of phase detector is also harmonic, for the impulse oscillations (for the same electronic realization of feedback loop) it is a continuous piecewise-linear periodic function.

In the present work the development of the above-mentioned technique for PLL is pursued. Here for the standard electronic realizations, the characteristics of phase detectors are computed and the differential equations, describing the PLL operation, are derived.

Here together with usual PLL the Costas loop is also considered. The essential conclusion is that the Costas loop with impulse oscillators tunes to a half frequency of master oscillator.

## 2 BLOCK DIAGRAM AND MATHEMATICAL MODEL OF PLL

Consider a PLL on the first level (Fig.1)

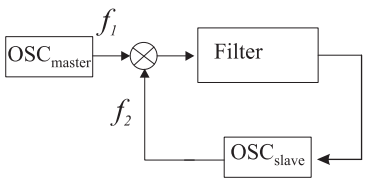



Figure 1: Electronic circuit of PLL.

Here OSC$_{\text{master}}$ is a master oscillator, OSC$_{\text{slave}}$ is a slave oscillator, which generates high-frequency "almost harmonic oscillations"

$$f_j(t) = A_j \sin(\omega_j(t)t + \psi_j). \qquad (1)$$

Block $\times$ is a multiplier of oscilations of $f_1(t)$ and $f_2(t)$. At its output the signal $f_1(t)f_2(t)$ arises. The relations between the input $\xi(t)$ and the output $\sigma(t)$ of

linear filter have the form

$$\sigma(t) = \alpha_0(t) + \int_0^t \gamma(t-\tau)\xi(\tau)\,d\tau.$$

Here $\gamma(t)$ is an impulse transient function of filter, $\alpha_0(t)$ is an exponentially damped function, depending on the initial date of filter at the moment $t = 0$.

Now we reformulate the high-frequency property of oscillations $f_j(t)$ to obtain the following condition.

Consider the great fixed time interval $[0,T]$, which can be partitioned into small intervals of the form $[\tau,\tau+\delta], (\tau \in [0,T])$, where the following relations

$$\begin{aligned} &|\gamma(t)-\gamma(\tau)| \le C\delta,\ |\omega_j(t)-\omega_j(\tau| \le C\delta, \\ &\forall t \in [\tau,\tau+\delta],\ \forall\tau \in [0,T], \end{aligned} \tag{2}$$

$$|\omega_1(\tau)-\omega_2(\tau)| \le C_1,\ \forall\tau \in [0,T], \tag{3}$$

$$\omega_j(t) \ge R,\ \forall t \in [0,T] \tag{4}$$

are satisfied. Here we assume that the quantity $\delta$ is sufficiently small with respect to the fixed numbers $T, C, C_1$, the number $R$ is sufficiently great with respect to the number $\delta$.

The latter means that on the small intervals $[\tau,\tau+\delta]$ the functions $\gamma(t)$ and $\omega_j(t)$ are "almost constants" and the functions $f_j(t)$ rapidly oscillate as harmonic functions. It is clear that such conditions occur for high-frequency oscillations.

Consider two block diagram described in Fig. 2 and 3.

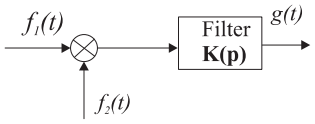



Figure 2: Multiplier and filter with transfer function $K(p)$.

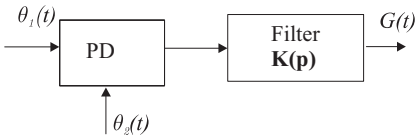



Figure 3: Phase detector and filter.

Here $\theta_j(t) = \omega_j(t)t + \Psi_j$ are phases of the oscillations $f_j(t)$, PD is a nonlinear block with the characteristic $\varphi(\theta)$, being called a phase detector (discriminator). The phases $\theta_j(t)$ enter the inputs of PD block and the output is the function $\varphi(\theta_1(t)-\theta_2(t))$.

The signals $f_1(t)f_2(t)$ and $\varphi(\theta_1(t)-\theta_2(t))$ enter the same filters with the same impulse transient function $\gamma(t)$. The filter outputs are the functions $g(t)$ and $G(t)$ respectively.

A classical PLL synthesis is based on the following result

*Theorem 1.* If conditions (2)–(4) are satisfied and

$$\varphi(\theta) = \frac{1}{2}A_1A_2\cos\theta,$$

then for the same initial data of filter the following relation

$$|G(t)-g(t)| \le C_2\delta,\ \forall t \in [0,T].$$

is valid. Here $C_2$ is a certain number not depending on $\delta$.

Thus, the outputs of two block-diagrams in Fig. 2 and Fig. 3: $g(t)$ and $G(t)$, differ little from each other and we can pass (from a standpoint of the asymptotic with respect to $\delta$) to the following description level, namely to the level of phase relations 2).

In this case a block diagram in Fig. 1 passes to the following block diagram (Fig. 4)

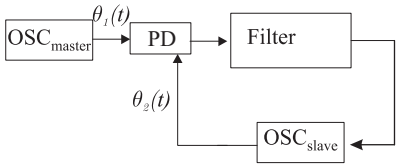



Figure 4: Block diagram of PLL on the level of phase relations.

Consider now the high-frequency oscillators, connected by a diagram in Fig. 1. Here

$$f_j(t) = A_j \text{sign}\,\sin(\omega_j(t)t + \psi_j). \tag{5}$$

We assume, as before, that conditions (2)–(4) are satisfied.

Consider a $2\pi$-periodic function $\varphi(\theta)$ of the form

$$\varphi(\theta) = \begin{cases} A_1A_2(1+2\theta/\pi) \text{ for } \theta \in [-\pi,0], \\ A_1A_2(1-2\theta/\pi) \text{ for } \theta \in [0,\pi]. \end{cases} \tag{6}$$

and block-diagrams in Fig. 2 and 3.

*Theorem 2.* If conditions (2)–(4) are satisfied and the characteristic of phase detector $\varphi(\theta)$ has the form (6), then for the same initial data of filter the following relation holds

$$|G(t)-g(t)| \le C_3\delta,\ \forall t \in [0,T].$$

Here $C_3$ is a certain number not depending on $\delta$.

Theorem 2 is a base for the synthesis of PLL with impulse oscillators. It permits us for the impulse clock oscillators to consider two block-diagrams in parallel: on the level of electronic realization (Fig. 1) and on the level of phase relations (Fig. 4), where the common principles of the phase synchronization theory can be applied. Thus, we can construct the theory

of phase synchronization for the distributed system of clocks in multiprocessor cluster.

Let us make a remark necessary to derive the differential equations of PLL.

Consider a quantity

$$\dot{\theta}_j(t) = \omega_j(t) + \dot{\omega}_j(t)t.$$

For the well-synthesized PLL, namely possessing the property of global stability, we have an exponential damping of the quantity $\dot{\omega}_j(t)$:

$$|\dot{\omega}_j(t)| \leq Ce^{-\alpha t}.$$

Here $C$ and $\alpha$ are certain positive numbers not depending on $t$. Therefore the quantity $\dot{\omega}_j(t)t$ is, as a rule, sufficiently small with respect to the number $R$ (see condition (2)– (4)).

From the above we can conclude that the following approximate relation

$$\dot{\theta}_j(t) = \omega_j(t) \tag{7}$$

is valid. When derived the differential equations of this PLL, we make use of a block diagram in Fig. 4 and relation (7),which is assumed to be valid precisely .

Note that, by assumption, the control low of tunable oscillators is linear:

$$\omega_2(t) = \omega_2(0) + LG(t). \tag{8}$$

Here $\omega_2(0)$ is the initial frequency of tunable oscillator, $L$ is a certain number, $G(t)$ is a control signal, which is a filter output (Fig. 4).

Thus, the equation of PLL is as follows

$$\dot{\theta}_2(t) = \omega_2(0) + L(\alpha_0(t) + \int_0^t \gamma(t-\tau).\, \varphi(\theta_1(\tau) - \theta_2(\tau))d\tau).$$

Assuming that the master oscillator such that $\omega_1(t) \equiv \omega_1(0)$, we obtain the following relations for PLL

$$\begin{aligned} &(\theta_1(t) - \theta_2(t))^{\bullet} + L(\alpha_0(t) + \int_0^t \gamma(t-\tau). \\ &\varphi(\theta_1(\tau) - \theta_2(\tau))d\tau) = \omega_1(0) - \omega_2(0). \end{aligned} \tag{9}$$

This is an equation of PLL.

Applying the similar approach, we can conclude that in PLL the filters with transfer functions of more general form can be used:

$$K(p) = a + W(p),$$

where $a$ is a certain number, $W(p)$ is a proper fractional rational function. In this case in place of equation (9) we have

$$\begin{aligned} &(\theta_1(t) - \theta_2(t))^{\bullet} + L(a(\varphi(\theta_1(t) - \theta_2(t)) + \\ &+ \alpha_0(t) + \int_0^t \gamma(t-\tau)\varphi(\theta_1(\tau) - \theta_2(\tau))d\tau) = \\ &= \omega_1(0) - \omega_2(0). \end{aligned} \tag{10}$$

In the case when the transfer function of the filter $a + W(p)$ is non degenerate, i.e. its numerator and denominator do not have common roots, equation (10) is equivalent to the following system of differential equations

$$\begin{aligned} \dot{z} &= Az + b\psi(\sigma) \\ \dot{\sigma} &= c^* z + \rho\psi(\sigma). \end{aligned} \tag{11}$$

Here $A$ is a constant $n \times n$-matrix, $b$ and $c$ are constant $n \times n$-vectors, $\rho$ is a number, $\psi(\sigma)$ is a $2\pi$-periodic function, satisfying the relations $\rho = -aL$

$$W(p) = L^{-1}c^*(A - pI)^{-1}b,$$
$$\psi(\sigma) = \varphi(\sigma) - \frac{\omega_1(0) - \omega_2(0)}{L(a + W(0))}.$$

Note that in (11) $\sigma = \theta_1 - \theta_2$.

Using Theorem 2, we can make the design of a block diagram of floating PLL, which plays a role of the function of frequency synthesizer and the function of correction of the clock-skew (see parameter $\tau$ in Fig. 5).

Such a block diagram is shown in Fig. 5.

Here $\text{OSC}_{\text{master}}$ is a master oscillator, *Delay* is a time-delay element, *Filter* is a filter with transfer function

$$W(p) = \frac{\beta}{p + \alpha},$$

$\text{OSC}_{\text{slave}}$ is a slave oscillator, PD1 and PD2 are programmable dividers of frequencies, *Processor* is a processor.

The *Relay* element plays a role of floating correcting block. The introduction of it allow us to null a residual clock skew, which arises for the nonzero initial difference of frequencies of master and slave oscillators.

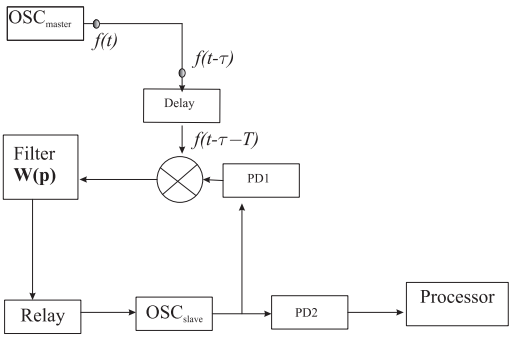



Figure 5: Block diagram of PLL.

Note, the electronic realization of clock and delay can be found in (Ugrumov, 2000; Razavi, 2003) and that of multipliers, filters, and relays in (Aleksenko, 2004; Razavi, 2003). The description of dividers of frequency can be found in (Solonina et al., 2000).

Assume, as usual, that the frequency of master oscillator is constant, namely $\omega_1(t) \equiv \omega_1 = const$. The parameter of delay line $T$ is chosen in such a way that $\omega_1(T+\tau) = 2\pi k + 3\pi/2$. Here $k$ is a certain natural number, $\omega_1\tau$ is a clock skew.

By Theorem 2 and the choice of $T$ the block diagram, shown in Fig. 6, can be changed by the close block diagram, shown in Fig. 6.

Here $2\pi$ is a periodic characteristic of phase detector. It has the form

$$\varphi(\theta) = \begin{cases} 2A_1A_2\theta/\pi & \text{for}\,\theta \in [-\frac{\pi}{2}, \frac{\pi}{2}] \\ 2A_1A_2(1-\theta/\pi) & \text{for}\,\theta \in [\frac{\pi}{2}, \frac{3\pi}{2}], \end{cases} \tag{12}$$

$\theta_2(t) = \theta_3(t)/M$, $\theta_4(t) = \theta_3(t)/N$, where the natural numbers $M$ and $N$ are the parameters of programmable divisions PD1 and PD2.

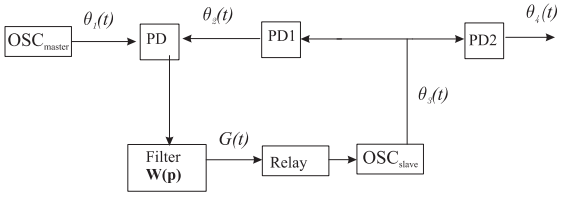



Figure 6: Equivalent block diagram of PLL.

For transient process (capture mode) the following conditions

$$\lim_{t\to+\infty}(\theta_4(t) - \frac{M}{N}\theta_1(t)) = \frac{2\pi kM}{N} \tag{13}$$

(phase capture)

$$\lim_{t\to+\infty}(\dot\theta_4(t) - \frac{M}{N}\dot\theta_1(t)) = 0 \tag{14}$$

(frequency capture) must be satisfied.

Relations (13) and (14) are the main requirements of PLL for array processors. The time of transient processors depends on the initial data and is sufficiently large for multiprocessors system (Leonov and Seledzhi, 2002; Kung, 1988). Here a difference between the beginning of transient process and the beginning of performance of parallel algorithm can be some minutes. This difference is very large for the electronic systems.

Assuming that the characteristic of relay is of the form $\Psi(G) = \text{sign}G$ and the actuating element of slave oscillator is linear, we have

$$\dot\theta_3(t) = R\text{sign}G(t) + \omega_3(0), \tag{15}$$

where $R$ is a certain number, $\omega_3(0)$ is the initial frequency, $\theta_3(t)$ is a phase of slave oscillator.

Taking into account relations (15), (1), (12), and the block diagram in Fig. 6, we have the following differential equations of PLL

$$\begin{aligned} &\dot G + \alpha G = \beta\varphi(\theta) \\ &\dot\theta = -\frac{R}{M}\text{sign}G + (\omega_1 - \frac{\omega_3(0)}{M}). \end{aligned} \tag{16}$$

Here $\theta(t) = \theta_1(t) - \theta_2(t)$.

# 3 CRITERION OF GLOBAL STABILITY OF PLL

System (16) can be written as

$$\begin{aligned} &\dot G = -\alpha G + \beta\varphi(\theta) \\ &\dot\theta = -F(G), \end{aligned} \tag{17}$$

where

$$F(G) = \frac{R}{M}\text{sign}G - (\omega_1 - \frac{\omega_3(0)}{M}).$$

*Theorem 3.* If the inequality

$$|R| > |M\omega_1 - \omega_3(0)| \tag{18}$$

is valid, then any solution of system (17) as $t \to +\infty$ tends to a certain equilibrium.

If the inequality

$$|R| < |M\omega_1 - \omega_3(0)| \tag{19}$$

is valid, then all the solutions of system (17) tends to infinity as $t \to +\infty$.

Consider the equilibria for system (17). For any equilibrium we have

$$\dot\theta(t) \equiv 0, \quad G(t) \equiv 0, \quad \theta(t) \equiv \pi k.$$

Theorem 4. Let relation (18) be valid. In this case, if $R > 0$, then the following equilibria

$$G(t) \equiv 0, \quad \theta(t) \equiv 2k\pi \tag{20}$$

are locally asymptotically stable and the following equilibria

$$G(t) \equiv 0, \quad \theta(t) \equiv (2k+1)\pi \tag{21}$$

are locally unstable. If $R < 0$, then equilibria (21) are locally asymptotically stable and equilibria (20) are locally unstable.

Thus, for relations (13) and (14) to be satisfied it is necessary to choice the parameters of system in such a way that the inequality holds

$$R > |M\omega_1 - \omega_3(0)|. \tag{22}$$

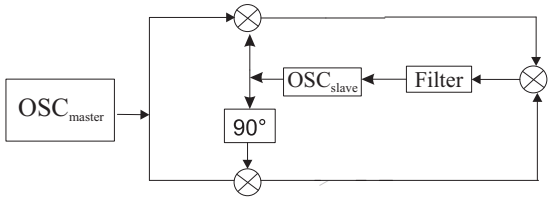



Figure 7: Costas loop.

# 4 COSTAS LOOP

Consider now a block diagram of the Costas loop (Fig. 7)

Here all denotations are the same as in Fig. 1, $90^o$ – is a quadrature component. As before, we consider here the case of the high-frequency harmonic and impulse signals $f_j(t)$.

However together with the assumption that conditions (2) and (4) are valid we assume also that (3) is satisfied for the signal of the type (1) and the relation

$$|\omega_1(\tau) - 2\omega_2(\tau)| \le C_1, \quad \forall \tau \in [0,T], \tag{23}$$

is valid for the signal of the type (5).

Applying the similar approach we can obtain differential equation for the Costas loop, where

$$\begin{aligned} \dot{z} &= Az + b\Psi(\sigma) \\ \dot{\sigma} &= c^* z + \rho\Psi(\sigma). \end{aligned} \tag{24}$$

Here $A$ is a constant $n \times n$-matrix, $b$ and $c$ are constant $n$-vectors, $\rho$ is a number, $\Psi(\sigma)$ is a $2\pi$-periodic function, satisfying the following relations

$$\rho = -2aL, \quad W(p) = (2L)^{-1} c^* (A - pI)^{-1} b,$$

$$\psi(\sigma) = \frac{1}{8} A_1^2 A_2^2 \sin\sigma - \frac{\omega_1(0) - \omega_2(0)}{L(a + W(0))},$$

$\sigma = 2\theta_1 - 2\theta_2$ (in the case (1);

$$\psi(\sigma) = P(\sigma) - \frac{\omega_1(0) - 2\omega_2(0)}{2L(a + W(0))},$$

$$P(\sigma) = \begin{cases} -2A_1^2 A_2^2 \left(1 + \dfrac{2\sigma}{\pi}\right), \ \sigma \in [0, \pi] \\ -2A_1^2 A_2^2 \left(1 - \dfrac{2\sigma}{\pi}\right), \ \sigma \in [-\pi, 0] \end{cases}$$

$\sigma = \theta_1 - 2\theta_2$ (in the case (5).

From the above equations it follows that for deterministic (when the noise is lacking) description of the Costas loops the conventional introduction of additional filters turns out unnecessary. Here a central filter plays their role.

# REFERENCES


Abramovitch, D., 2002. *Phase-Locked Loope A control Centric. Tutorial*, in the Proceedings of the 2002 ACC.

Aleksenko, A., 2004. *Digital engineering*, Unimedstyle. Moscow. (in Russian)

Best Ronald E., 2003. *Phase-Lock Loops: Design, Simulation and Application*, McGraw Hill, 5ed.

Egan, W.F., 2000. *Frequency Synthesis by Phase Lock* ,(2nd ed.), John Wiley and Sons, 2ed.

Gardner F., 2005. *Phase–lock techniques*, John Wiley & Sons, New York, 2ed.

Kroupa, V., 2003. *Phase Lock Loops and Frequency Synthesis*, John Wiley & Sons.

Kung, S., 1988. *VLSI Array Processors*, Prentice Hall. New York.

Lapsley, P., Bier, J., Shoham, A., Lee, E., 1997. *DSP Processor Fundamentals Architecture and Features*, IEE Press. New York.

Leonov, G., Reitmann, V., Smirnova, V., 1992. *Nonlocal Methods for Pendulum-Like Feedback Systems*, Teubner Verlagsgesselschaft. Stuttgart; Leipzig.

Leonov, G., Ponomarenko, D., Smirnova, V., 1996. *Frequency-Domain Methods for Nonlinear Analysis. Theory and Applications*, World Scientific. Singapore.

Leonov, G., Seledzhi, S., 2002. *Phase locked loops in array processors*, Nevsky dialekt. St.Petersburg. (in Russian)

Lindsey, W., 1972, *Sinchronization systems in communication and ontrol*, Prentice-Hall. New Jersey.

Lindsey, W., Chie, C., 1981. A Survey of Digital Phase Locked Loops. *Proceedings of the IEEE*.

Razavi, B., 2003. *Phase-Locking in High-Performance Systems: From Devices to Architectures*, John Wiley & Sons.

Smith, S., 1999. *The Scientist and Engineers Guide to Digital Dignal Processing*, California Technical Publishing. San Diego.

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.

# DISTURBANCES ESTIMATION FOR MOLD LEVEL CONTROL IN THE CONTINUOUS CASTING PROCESS

Karim Jabri[1,2], Bertrand Bele[1], Alain Mouchette[1]
[1] *Measurement Control Engineering Department, ArcelorMittal Research, Maizières-Lès-Metz, France*
*karim.jabri@arcelormittal.com, bertrand.bele@arcelormittal.com, alain.mouchette@arcelormittal.com*

Emmanuel Godoy [2], Didier Dimur[2]
[2] *Department of Automatic Control, Supélec,Gif sur Yvette, France*
*emmanuel.godoy@supelec.fr, didier.dumur@supelec.fr*

Keywords: Disturbance estimation, Luenberger observer, Harmonic disturbance, Continuous casting, Mold level control.

Abstract: This paper addresses the problem of mold level fluctuations in the continuous casting process, which strongly penalize the quality of the final product and lead to a costly machine downtime. Therefore, the mold level is controlled using a stopper as the flow control actuator and a level sensor. Under normal casting conditions, the current controllers provide suitable performances but abnormal conditions require manual intervention, such as the decrease of the casting speed, in particular when undesired disturbances like clogging/unclogging or bulging occur. These disturbances increase in severity for certain steel grades or at high casting speeds. Therefore, this paper focuses on the on-line disturbances estimation in order to introduce compensation actions. Starting with the presentation of the continuous casting process, the description of the model of the machine, and highlighting the main control challenges, an observer estimating clogging and bulging disturbances is then developed. This design may help future control architectures based on disturbances estimation. The proposed observer is finally validated by extracting disturbances from experimental signals measured on a continuous casting plant.

## 1 INTRODUCTION

Nowadays more than 96% of steel is produced by means of a continuous casting process which has significantly improved plant productivity in comparison with other solidification processes. Accurate control of the molten steel level in the mold is an important task from both the operating and quality points of view. Indeed, on the one hand, it is important to control the mold level to avoid molten steel overflows or mold emptying. On the other hand, the mold level must be kept constant to avoid alumina inclusions and slag being caught up in the molten steel, leading to defects associated with cracks in the slabs.

Under normal casting conditions, currently implemented controllers provide suitable performances. However, more severe operating conditions still require manual intervention, such as the decrease of the casting speed, in particular when undesired clogging/unclogging or bulging disturbances occur. These disturbances are in particular extremely sensitive for certain steel grades or at high casting speeds. Therefore, on-line disturbance estimations become an important challenge with a view to introducing feedforward actions within the control law.

The paper is structured as follows. Section 2 describes the continuous casting process and the model of the machine. The design of the observer is presented in Section 3, successively estimating clogging, bulging and both disturbances. The proposed observer is finally validated in Section 4 by estimating disturbances from experimental signals measured on a continuous casting plant.

## 2 CONTINUOUS CASTING PLANT MODEL

In the continuous casting process, as shown in Figure 1, molten steel flows from the tundish into the mold through the nozzle where it freezes against water-cooled mold walls to form a solid shell. There are several rolls below the mold to withdraw the solidified steel continuously from the bottom of the mold. The mold has an oscillatory movement with a magnitude of a few millimetres and a frequency of about 2 Hz that makes shell extraction easier. At the outlet of the machine, the steel is fully solidified and is cut into slabs.

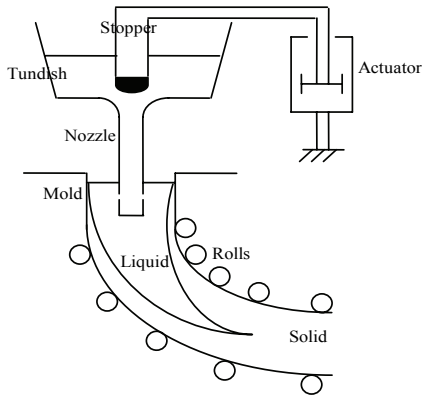



Figure 1: Continuous casting machine.

Usually, the casting speed is kept constant. The flow out of the mold is thus constant and only the inflow is controlled.

Figure 2 shows the major components of the mold level control model without any disturbance that is usually considered in the plants for the design of the control law,

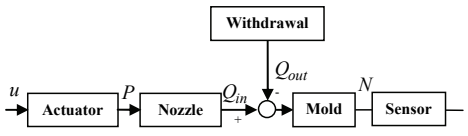



Figure 2: Plant model.

where $u$ is the control input, $P$ the stopper position, $N$ the mold level, $Q_{in}$ and $Q_{out}$ the steel flow-rate into and out of the mold.

This behavioral model shows that the flow into the mold $Q_{in}$ is regulated by the stopper position $P$ by acting on the control input $u$ of a hydraulic actuator. The nozzle is usually modelled in most cases by a simple gain. The flow out of the mold $Q_{out}$ is imposed by the casting speed. The mold level $N$ is thus given by the integration of the difference between $Q_{in}$ and $Q_{out}$ divided by the cross section of the mold. This level is measured by a sensor which can be either an eddy current or a floating one, returning only local level and not the whole free surface feature. Transfer functions appearing in the plant model are summarized in Table 1,

Table 1: Transfer functions of the plant model.

| Block | Transfer function |
|---|---|
| Actuator | $\frac{G_a}{s(1+\tau_1 s)}$ |
| Nozzle | $G_s$ |
| Withdrawal | $S\ v$ |
| Mold | $\frac{1}{S\ s}$ |
| Level sensor | $\frac{G_{ss}}{1+\tau_2 s}$ |

where $G_s$ is the stopper gain, $G_{ss}$ the level sensor gain, $G_a$ the actuator gain, $S$ the mold section, $\tau_n$ the nozzle delay, $\tau_1$ the actuator time constant, $\tau_2$ the time constant of the level sensor, $v$ the casting speed and $s$ the Laplace variable.

The control objective is to maintain the mold level at a specified constant setpoint while limiting the level fluctuations as much as possible. Implemented controllers use both the level and stopper signals as available measurements and elaborate the actuator input $u$ acting as the manipulated variable. Classic control structures currently working on real plants are of two types: a first structure considers regulation of the mold level with a PID controller; a second one includes two cascaded loops, regulating the stopper position in an inner loop through a proportional gain and the mold level in an external loop by means of a PI controller.

However, the control becomes more complex when disturbances occur on the plant or when the operating conditions change, leading to an unstable behavior. In fact, several phenomena disturb the balance between the flow into and out of the mold, causing fluctuations over the meniscus surface or an abrupt increase of the mold level. The standard controllers are not designed for such casting conditions. Therefore, new control strategies should be designed, for example those based on disturbances rejection. For this purpose, the disturbances must first be estimated on-line so that the control structure can compensate their influence

(Furtmueller and Gruenbacher, 2006). Section 3 will thus consider the design of observers able to estimate the two most important disturbances occurring during continuous casting: the clogging/unclogging and the bulging.

# 3 DISTURBANCE ESTIMATION

This Section will successively consider observers for clogging/unclogging disturbance, then bulging disturbance and finally for both types.

## 3.1 Clogging

The clogging event is one of the most serious phenomena faced by the operators in the continuous casting machine. It increases operating cost and decreases productivity. Clogging takes place essentially at the nozzle wall even if in principle it can occur anywhere inside the nozzle. Nozzle clogging takes many different forms. The first one is the sediment of solid inclusions already present in the steel entering the nozzle. The second form is related to air aspiration into the nozzle through joints which leads to reoxidation. The third type of clogging is attributed to reactions between aluminum in the steel and an oxygen source in the refractory. In practice, a given nozzle clog is often a combination of several of these types (Thomas and Bai, 2001). The clogging effect is not an instantaneous phenomenon but develops over time. Its cycle consists of a phase of slow clogging, followed by a sudden unclogging that raises considerably the mold level. Its period is random (Dussud *et al*., 1998).

This part sets out to propose a clogging estimation procedure assuming that this is the only disturbance to arise during the casting operations. Currently, when the clogging occurs, the controller forces the stopper to move to maintain a constant casting speed and a constant level inside the mold. When the unclogging occurs, the flow passage inside the nozzle becomes larger. The stopper position decreases suddenly to reduce the control zone (opening surface) in order to maintain a constant flow rate into the mold. During the clogging/unclogging cycle, the stopper position is thus a succession of ramps. Therefore, a possible model of the clogging/unclogging phenomenon can be the inclusion of an additional flow $d_{\text{clog}}$ to $Q_{in\_ideal}$ (which is the real inflow without clogging), as in Figure 3. $d_{\text{clog}}$ is also a succession of ramps because the flow-rate out of the mold $Q_{out}$ and the level $N$ are generally constant. $d_{\text{clog}}$ has thus the same behavior as the stopper position $P$.

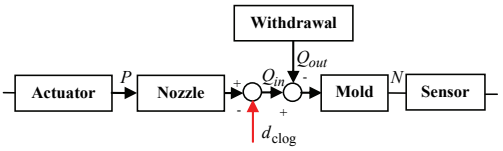



Figure 3: Plant model taking clogging into account.

The observer is designed considering the stopper position $P$ and the casting speed $v$ as inputs, and the mold level $N$ as output. In the following modelling phase, the time constant of the level sensor is ignored, since it influences frequencies out of the bandwidth of the regulation. Without loss of generality, $G_{ss}$ is assumed to be equal to 1. According to the previous figure:

$$S\dot{N} = \underbrace{G_s P}_{Q_{in\_ideal}} - \underbrace{S v}_{Q_{out}} - d_{\text{clog}} \quad \text{with } \ddot{d}_{\text{clog}} = 0 \tag{1}$$

Thus, the clogging model under the state space formalism is given by:

$$\begin{cases} \dot{X}_{\text{clog}} = A_{\text{clog}} X_{\text{clog}} + B_{\text{clog}} U_{c\,\text{log}} \\ N = C_{\text{clog}} X_{\text{clog}} \end{cases} \tag{2}$$

with:

$$X_{\text{clog}} = \begin{pmatrix} N \\ d_{\text{clog}} \\ \dot{d}_{\text{clog}} \end{pmatrix},\ U_{\text{clog}} = \begin{pmatrix} P \\ v \end{pmatrix},\ C_{\text{clog}}^T = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$$
$$A_{\text{clog}} = \begin{pmatrix} 0 & -\frac{1}{S} & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix},\ B_{\text{clog}} = \begin{pmatrix} \frac{G_s}{S} & -1 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} \tag{3}$$

The observability matrix $O_{\text{clog}}$ has full rank 3. The system is therefore completely observable. Based on the model Eq. 2, the Luenberger observer (Sontag, 1998) is given as follows:

$$\dot{\hat{X}}_{\text{clog}} = (A_{\text{clog}} - K_{\text{clog}} C_{\text{clog}})\hat{X}_{\text{clog}} + B_{\text{clog}} U_{\text{clog}} + K_{\text{clog}} N \tag{4}$$

where $K_{\text{clog}}$ is the observer gain chosen so that the observer is stable and achieves a desired dynamic. The observer converges if the eigenvalues of the square matrix $A_{\text{clog}} - K_{\text{clog}} C_{\text{clog}}$ are strictly negative. $K_{\text{clog}}$ is in a first step chosen to satisfy this condition. The observer is afterwards adjusted with

dynamics as fast as possible, the compromise being that its stability decreases with increasing dynamics.

## 3.2 Bulging

During the continuous casting process of a slab, the volume of liquid steel inside the solidified shell can be changed by strand bulging in the secondary cooling zone. The bulging occurs between rolls due to increasing pressure inside the strand. It is divided into static or dynamic bulging, according to the strand movement, and steady or unsteady bulging according to the variation with time (Yoon *et al.*, 2002). The most disruptive type is the unsteady bulging generating level fluctuations in the mold.

This part sets out to propose a bulging estimation procedure assuming that this is the only disturbance to arise during the casting operations. It is supposed that the bulging profile at each site between two rolls is described by a sine function (Lee and Yim, 2000) with a frequency between 0.05 and 0.15 Hz. Therefore, this displacement induces changes in the flow-rate out of the mold. The bulging phenomenon can thus be modelled by an additional flow $d_{\text{bulge}}$ to $Q_{out\_ideal}$ (which is the real outflow without bulging). $d_{\text{bulge}}$ is a sum of several sine waves. To determine its frequencies, the level signal spectrum must be calculated and the most significant frequencies belonging to the frequency range selected. In the following part of this subsection, and without loss of generality, only two frequencies of $d_{\text{bulge}}$ are considered (see Figure 4).

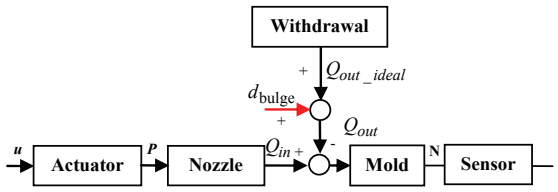



Figure 4: Plant model taking bulging into account.

For the following modelling phase, the time constant of the level sensor is again ignored. According to Figure 4:

$$S\dot{N} = \underbrace{G_S\ P}_{Q_{in\_ideal}} - \underbrace{(S\ v + d_{\text{bulge}\,1} + d_{\text{bulge}\,2})}_{Q_{out}} \quad (5)$$

$$\text{with } d_{\text{bulge}\ i} = A_i \cos(\omega_i t + \varphi_i)\ \ i = 1,2$$

The bulging model under the state space formalism is thus given by:

$$\begin{cases} \dot{X}_{\text{bulge}} = A_{\text{bulge}} X_{\text{bulge}} + B_{\text{bulge}} U_{\text{bulge}} \\ N = C_{\text{bulge}} X_{\text{bulge}} \end{cases} \quad (6)$$

$$\text{with } X_{\text{bulge}} = \begin{pmatrix} N \\ d_{\text{bulge}1} \\ \dot{d}_{\text{bulge}1} \\ d_{\text{bulge}2} \\ \dot{d}_{\text{bulge}2} \end{pmatrix},\ U_{\text{bulge}} = \begin{pmatrix} P \\ v \end{pmatrix}$$

$$A_{\text{bulge}} = \begin{pmatrix} 0 & -\frac{1}{S} & 0 & -\frac{1}{S} & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & -\omega_1^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & -\omega_2^2 & 0 \end{pmatrix}, \quad (7)$$

$$B_{\text{bulge}} = \begin{pmatrix} \frac{G_s}{S} & -1 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix},\ C_{\text{bulge}}^T = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

The observability $O_{\text{bulge}}$ has full rank 5. The system is thus completely observable. Based on the model Eq. 6, the Luenberger observer is given by the following equation:

$$\dot{\hat{X}}_{\text{bulge}} = (A_{\text{bulge}} - K_{\text{bulge}} C_{\text{bulge}})\hat{X}_{\text{bulge}} + \\ + B_{\text{bulge}} U_{\text{bulge}} + K_{\text{bulge}} N \quad (8)$$

where $K_{\text{bulge}}$ is the observer gain adjusted as in Section 3.1.

## 3.3 Clogging and Bulging

The previous clogging and bulging observers were designed separately to estimate respectively clogging or bulging being the only disturbance acting on the system. However, when clogging and bulging occur simultaneously during the continuous casting operation, these two observers must be merged into a single global one that will be able to estimate $d_{\text{clog}}$ and $d_{\text{bulge}}$ individually (Figure 5).

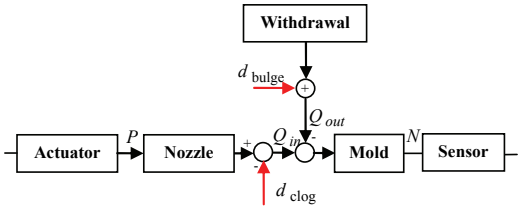



Figure 5: Plant model with clogging and bulging effects.

With this figure, merging Eqs. 1 and 5 leads to:

$$\begin{aligned} S\dot{N} &= \underbrace{G_s\ P - d_{\text{clog}}}_{Q_{in}} - \underbrace{(S\ v + d_{\text{bulge}})}_{Q_{out}} \\ \text{with}\ \ \ddot{d}_{\text{clog}} &= 0 \\ d_{\text{bulge}} &= d_{\text{bulge}1} + d_{\text{bulge}2} \\ d_{\text{bulge}\ i} &= A_i \cos(\omega_i t + \varphi_i)\ \ i = 1,2 \end{aligned} \tag{9}$$

which results in the global clogging/bulging model under the state space formalism:

$$\begin{cases} \dot{X}_{est} = A_{est} X_{est} + B_{est} U_{est} \\ N = C_{est} X_{est} \end{cases} \tag{10}$$

with:

$$X_{est}^T = \begin{pmatrix} N & d_{\text{clog}} & \dot{d}_{\text{clog}} & d_{\text{bulge}1} \\ & & & \dot{d}_{\text{bulge}1}\ d_{\text{bulge}2}\ \dot{d}_{\text{bulge}2} \end{pmatrix}$$

$$A_{est} = \begin{pmatrix} 0 & -\frac{1}{S} & 0 & -\frac{1}{S} & 0 & -\frac{1}{S} & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -\omega_1^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & -\omega_2^2 & 0 \end{pmatrix} \tag{11}$$

$$U_{est} = \begin{pmatrix} P \\ v \end{pmatrix},\ B_{est} = \begin{pmatrix} \frac{G_s}{S} & -1 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix},\ C_{est}^T = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

The observability matrix $O_{est}$ has full rank 7. The system is thus completely observable. Based on the model Eq. 10, the global Luenberger observer is given by the following equation, where $K_{est}$ is the observer gain adjusted as previously:

$$\begin{aligned} \dot{\hat{X}}_{est} &= (A_{est} - K_{est} C_{est})\hat{X}_{est} + \\ &+ B_{est} U_{est} + K_{est} N \end{aligned} \tag{12}$$

The generalization to more than two frequencies taken into account in the bulging disturbance is performed by considering two additional state variables per added frequency, with adequate rows and columns in the state representation. It must be noticed that, due to the observer structure, the average of the $d_{\text{bulge}}$ signal is included in the estimation of $d_{\text{clog}}$. In short, $d_{\text{clog}}$ contains the average of $d_{\text{bulge}}$ and a succession of ramps possibly.

# 4 OBSERVER VALIDATION

The previous global observer is now applied on experimental data registered during continuous casting operations. A first experiment is considered (record 1), for which the eigenvalues of the observer have been tuned (according to bulging signal frequencies) to −1.5, −1.57, −1.27, −1.42, −1.35, −1.2 and −1.12. Results are given in Figures 6 and 7. Figure 6 compares the estimated $d_{\text{bulge}}$ to $Q_{out\_ideal}$. $Q_{out\_ideal}$ appears to be ten times greater than $d_{\text{bulge}}$. It can be concluded that there is no bulging effect in this first record.

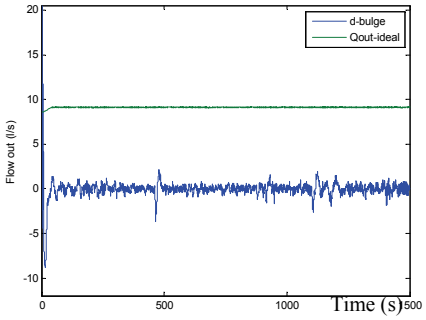



Figure 6: Comparison between the estimated $d_{\text{bulge}}$ and $Q_{out\_ideal}$ in the case of the first record.

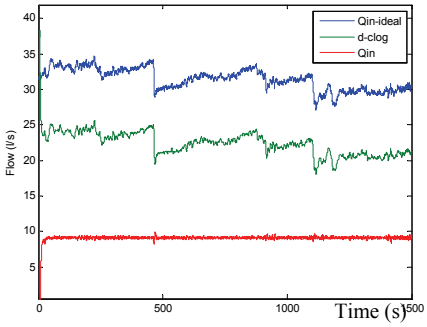



Figure 7: Comparison between the estimated $d_{\text{clog}}$, the ideal and the real input into the mold for the first record.

Knowing this, Figure 7 now compares the estimated $d_{\text{clog}}$ elaborated by the observer with the ideal ($Q_{in\_ideal}$) and the real ($Q_{in}$) input into the mold in the case of this first record. $Q_{in\_ideal}$ is recomputed by means of the measurement of the stopper position $P$ as mentioned in Eq. 5 and $Q_{in}$ is recalculated by means of the relation in Eq. 9. It is shown that the estimated clogging disturbance follows the expected profile, i.e. ramp variations during the clogging phase and a sudden decrease due to unclogging. This figure also illustrates that $Q_{in}$ is three times smaller than the ideal value $Q_{in\_ideal}$

expected without clogging. Thus this estimation by means of an observer, which may be useful for control purposes, also helps to quantify the intensity of the clogging phenomenon.

The observer, tuned as previously, is now applied to a second experiment (record 2). In this case, three frequencies of the signal $d_{\text{bulge}}$ have to be considered, respectively 0.084, 0.095 and 0.082 Hz as the most significant in the bulging frequency range [0.05, 0.15] Hz.

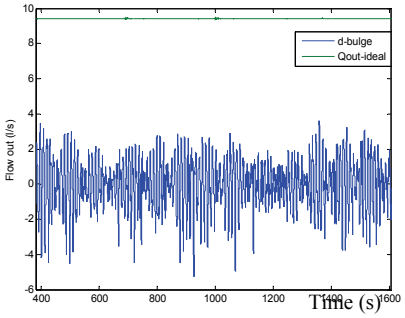



Figure 8: Comparison between the estimate $d_{\text{bulge}}$ and $Q_{out\_ideal}$ in the case of the second record.

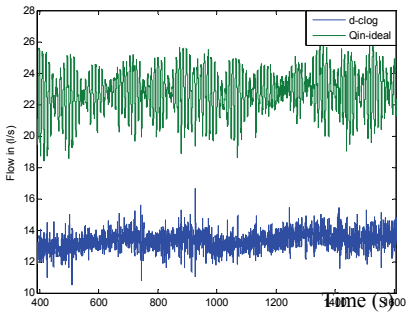



Figure 9: Comparison between the estimated $d_{\text{clog}}$ and $Q_{in\_ideal}$ in the case of the second record.

Figures 8 and 9 show the estimated disturbances $d_{\text{bulge}}$ and $d_{\text{clog}}$. Figure 9 compares the estimation of $d_{\text{clog}}$ to $Q_{in\_ideal}$. Without ramp in $d_{\text{clog}}$, it can be concluded that $d_{\text{clog}}$ only represents the average of the $d_{\text{bulge}}$ signal of Figure 8 which should be added to it in order to estimate its correct value, and that there is no clogging effect in this record. From a control point of view as described in (Furtmueller *et al.*, 2005), to have the average of $d_{\text{bulge}}$ included in $d_{\text{clog}}$ is not problematic. In fact, the control structure which compensates disturbances should use the sum of $d_{\text{clog}}$ and $d_{\text{bulge}}$ as an input and not the two disturbance estimations separately.

# 5 CONCLUSIONS

This paper presents the elaboration of a global observer designed to estimate clogging and bulging disturbances appearing in a continuous casting process. These estimations may be further used as inputs to compensation modules within mold level control structures. This observer is built with behavioral models of the physical process, assuming that these disturbances can be modelled as exogenous signals. Further research may consider a nonlinear nozzle gain to model the clogging effect, robustness analysis of the estimator, particularly with respect to variations of the bulging signal frequencies and model uncertainties.

# REFERENCES

Thomas, B.G., Bai, H., 2001. Tundish nozzle clogging – application of computational models. In *18rd PTD Conf. Proc.*, Baltimore (US).

Sontag, E., 1998. *Mathematical Control Theory: Deterministic Finite Dimensional Systems*, Springer. 2nd edition.

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.

Lee, J.D., Yim, C.H., 2000. The mechanism of unsteady bulging and its analysis with the finite element method for continuously cast steel. ISIJ International, 40(8):765-770.

Dussud, M., Galichet, S., Foulloy L.P., 1998. Application of fuzzy logic control for continuous casting mold level control. *IEEE Transactions on Control Systems Technology*, 6(2).

Furtmueller, C., Gruenbacher, E., 2006. Suppression of periodic disturbances in continuous casting using an internal model predictor. In *Proc. IEEE Intl. Conf. on Control Applications*, Munich, Germany.

Furtmueller, C., Del Re, L., Bramerdorfer, H., Moerwald, K., 2005. Periodic disturbance suppression in a steel plant with unstable internal feedback and delay. In *Proc. 5th Intl. Conf. on Tech. and Automation*, Greece.

# A PROTOTYPE FOR ON-LINE MONITORING AND CONTROL OF ENERGY PERFORMANCE FOR RENEWABLE ENERGY BUILDINGS

Benjamin Paris, Julien Eynard, Gregory François, Thierry Talbert and Monique Polit
*Laboratoire ELIAUS, Université de Perpignan Via Domitia, 52 avenue Paul Alduy, 66860 Perpignan, France*
*{benjamin.paris, julien.eynard, gregory.francois, talbert, polit}@univ-perp.fr*

Keywords: Renewable energies, Energy performance indicators, Monitoring system, Smart transducers, Control algorithms, Predictive control, Optimal control, High efficiency buildings.

Abstract: In this article, ways for improving the energetic performance of buildings are investigated. A state of the art leads to the introduction of a performance indicator expressed in kWh/m$^2$/yr. To improve the value of this indicator, a processor-based prototype of a real-time data-acquisition and monitoring system is developed in collaboration with two industrial companies. The set of measurements and corresponding sensors that are necessary to compute the value of the indicator while being consistent with the natural segmentation of energy consumption, is listed, thanks to the representation of the building using a systemic approach. Control algorithms are tested in simulation to improve renewable energy consumption while reducing fossil energy dependence, which are deemed to be applicable in practice using the proposed electronics. Simulations concerning the control and optimization of the power applied to two warmers in a room show large potential for fossil energy consumption reduction.

## 1 INTRODUCTION

Nowadays, it is widely admitted that climate change is induced by the intense human activity, and that greenhouse effect gases (GEG) exhaustion is one of the main contributors to this phenomenon. Hence, the decision to stabilize or to reduce GEG emission was taken in the late nineties by most of the industrialized country.

In France, 25% of GEG emissions and 46% of global energy consumption (Ademe, 2007) are due to the buildings. Using legal documentation, e.g. “Réglementation Thermique 2005” (RT2005), or “Diagnostic de Performance Energétique” (DPE), (Sesolis, 2006), French government would like to restrict building energy consumption while limiting wastefulness. Labels are investigated to promote good practice and make the French public opinion sensitive to these issues. In Europe, the situation is similar, witness the development of Swiss and German labels: “Minergie” and “Passivhause”, respectively. Hence, performance of building materials, design or management, needs to be improved.

However, one of the main difficulties when trying to achieve this purpose lies in the fact that energy consumption may vary from a building to another. In addition, energy consumption is segmented in terms of objectives. In this context, the method of choice for improving building energetic behaviour without reducing comfort is obviously to reduce the dependency to fossil energy by, e.g., developing the use of renewable energies. To achieve this goal, it is needed to: (i) characterize global and segmented energy consumption in a building, (ii) compute a performance indicator that takes into account the environment of the building, as well as the way energy is consumed, (iii) acquire and process data measurements to monitor energy consumption, (iv) propose control and optimization strategies for promoting the use of renewable energies.

The goal of this work, performed in collaboration with Apex BP Solar, Pyrescom and CSTB (Centre Scientifique et Technique du Bâtiment), is to develop a prototype of a commercially viable tool that will be able to perform the four aforementioned tasks. To be cost-effective, the tool needs to be small and easy to handle, to remain relatively cheap, to avoid the implementation of many sensors, to be applicable to

various buildings, regardless their localization, and to propose solutions depending on energy consumption segmentation.

In this context, the goal of this paper is not only to discuss independently the choice of the electronics or the choice of a specific control law but rather to present the approach as a whole. To estimate energetic performance, an indicator is necessary that is firstly defined. To compute this indicator, the set of needed measurements and corresponding sensors is listed. These sensors are also capable of providing information about the segmentation of energy consumption. To process the acquired data, appropriate electronics are needed. Hence, a processor-based electronic architecture is proposed. The advantages of the use of a processor instead of a standard microcontroller are discussed. Finally, control laws should be implemented to reduce fossil energy consumption. Hence, such laws (potentially applicable using the chosen processor) are investigated in simulation to enforce the use of renewable energies.

The corresponding simulated illustrative example deals with the energy consumption reduction in a room, which is assumed to be equipped with two controllable warmers, respectively using renewable ($W_{RE}$) and fossil ($W_F$) energies. To reduce fossil energy consumption, a mix of on-line and predictive control laws is proposed and compared to open loop simulations and to standard online control laws. The general underlying idea is to use $W_F$ if predictions or measurements indicate that $W_{RE}$ reaches saturation. In simulation, this approach leads to a large energy consumption reduction.

This article is organized as follows: Section 2 discusses the choice of a performance indicator. Section 3 presents the prototype of the data-acquisition system, while its applicability for on-line control and optimization of the energetic behaviour of buildings is investigated in simulation in Section 4. Finally, Section 5 concludes the paper.

# 2 PERFORMANCE INDICATOR

## 2.1 Choice of the Indicator

Almost twenty years ago, the energetic performance of building was not a strong preoccupation for governments, building material suppliers or real estate developers. Then, energy performance turned to a priority due to the impact of greenhouse effect gases together with the high level of energy costs. The first indicator proposed was a measurement of energy consumption (Duffaure-Gallais, 2006). However, it did not allow performing comparisons regarding localization or areas of the buildings. Recent researches provided specific documentation, which explains the method for computing a global indicator, i.e. annual energy consumption per square meter ($kWh/m^2/yr$), and fixes clear objectives in terms of energy consumption. This unit allows comparison between different buildings, with different constraints.

In France, successive governments have been showing a strong will to reduce human impact on climate (Maïzi, 2007), witness the attribution of "HPE" and "THPE" (Journal Officiel, 2006) labels whenever the energetic performance is 10% or 20% less than standard energy consumption. The underlying idea is very similar to the American "Energy Star" (Boyd, 2007) that is used in industry. Recently, the "HPE ENR" label was created to promote renewable energies. For old constructions, the DPE (Energetic Performance Diagnosis) label is expressed in $kWh/m^2/yr$ as well. Software, such as "3CL Excel©" (CSTB), which are based on building materials parameters (thermal conductivity insulation, glazing losses …), on the building design or equipment, can be used to compute the DPE index to classify buildings according to their levels of energy consumption.

However, the chosen indicator only provides a cumulated indication about energy consumption that aggregates different consumptions, e.g. for heating or cooling. In practice, European labels do not always use the same variables to compute this indicator and do not fix the same goals to reach: "Minergie" label suggests 42 $kWh/m^2/yr$ only for heating, while "PassivHause" label considers 30 $kWh/m^2/yr$ as normal energy consumption for heating and ventilation...

## 2.2 Environmental Factors and Energy Segmentation

In order to establish a fair diagnosis of the energetic behaviour of a building and to control energy consumption, buildings can be represented as dynamic systems, interacting with their environment, which consume energy with regard to different objectives (fig. 1).

It is proposed to focus on the following environmental factors:

1. Indoor and outdoor temperatures that can be acquired with smart transducers…

2. Wind and solar radiation that can help explaining many consumption levels or provide information about renewable energy availability.
3. Indoor relative humidity, which represents a specific comfort parameter and, thus, affects the user's behaviour.
4. If meteorological predictions are available, which is recommended, pressure can also be measured.

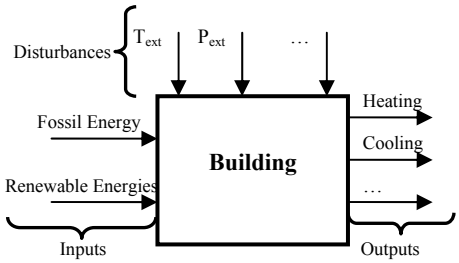



Figure 1: Building System and Interactions.

Buildings use different energies with a large emphasis on electricity. In the case of several kinds of sources (fuel, gas, electricity are used), computation rules exist to estimate the individual contributions to the indicator. The following list summarizes the main sources (inputs) of energy consumed in buildings, together with the objectives (outputs):

1. Electricity (includes heating …).
2. Specific electricity: (electricity that cannot be substituted by any other type of energy).
3. Cooling and ventilation energy.
4. Heating energy (apart from electricity): fuel oil, gas or wood.

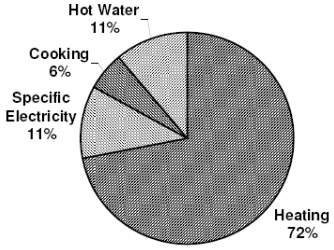



Figure 2: Segmentation of Energy End-Uses in Household (Ademe).

Figure 2 provides the typical energy consumption segmentation. Obviously, the main output is heating, hence the need for focusing on heating control and optimization for reducing energy consumption.

# 3 INSTRUMENTATION AND DATA ACQUISITION

## 3.1 Monitoring System Prototype

The acquisition of the aforementioned variables requires the choice of appropriate electronics. However: (i) implementation should be easy and (ii) total cost should remain rather low. In collaboration with our industrial partners, such architecture was developed and implemented at three different locations (Apex BP Solar and Pyrescom headquarters and at the University of Perpignan).

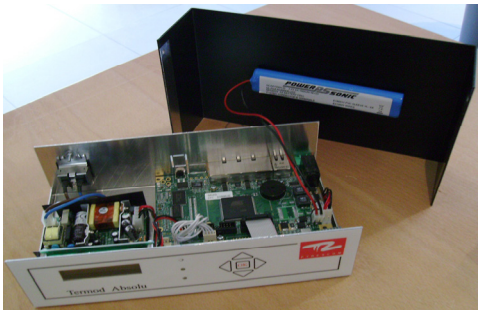


Figure 3: Monitoring System Prototype.

The prototype, which can be seen on Figure 3, is divided into two separate parts: (i) a core-bloc (composed of a low power processor, corresponding memory, and integrated hosts controllers), and (ii) a set of adaptable bloc sensors, which means that it is possible to record and process different data.

## 3.2 Data Acquisition System

As mentioned, with the chosen architecture, it is possible to use both information concerning energy consumption segmentation and operating conditions measurements. The smart transducers transmit data to the monitoring system discussed in the next subsection, through preferably a Controlled Area Network (CAN) bus. To avoid drilling, or pulling cable, wireless or Power Line Communication (PLC) systems are also studied.

Constraints discussed previously have been taken into account and the quantity and localization of implemented transducers depend on the interactions within the building and on the impact of disturbances.

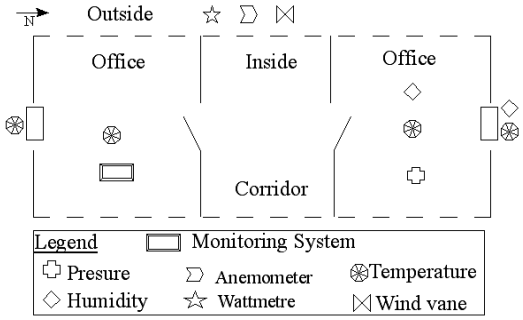



Figure 4: Smart Transducers and Monitoring System Localization (University Offices).

Figure 4 shows the example of the University's offices, where one of the three experimental setups is under implementation. Researches (Hensen, 1991) on heating control systems or on energetic efficiency (Mendoça, 2003, 2007) showed that north front temperature and inside temperature measurements are definitely musts for heating control purposes.

In addition, a compromise was found between the number of transducers, avoiding information redundancy and total cost. To generalize this approach to a broader range of customers, indirect measurements were preferred whenever possible. Note that, for confidentiality reasons, more details concerning the sensors cannot be given. However, all these sensors can interact with the processor described in the following subsection.

## 3.3 Core Architecture

It is proposed to use an ARM9© processor instead of a microcontroller, which is typically used in the metrology literature (Gungor, 1997, Leong, 1998), since:

1. ARM9© has a low level of energy consumption.
2. Hosts controllers are already integrated for: (i) connectivity, (ii) control purposes, (iii) human interface (CSI, Keypad…), (iv) memory expansion (MMC, PCMCIA…), and (v) providing e.g. Bluetooth communication.
3. Computation power is higher (4-8 bits versus 32-64 bits, 40 MHz versus 100-400 MHz).
4. Its high level of memory allows the handling of a higher number of different kinds of signals (Segars, 1998, Xingwu, 2006).
5. Control laws can be implemented, e.g. energy consumption prediction (Kalogirou, 2000), fuzzy logic (Lygouras, 2006) or fault diagnosis (Kalogirou, 2007).

# 4 ILLUSTRATIVE EXAMPLE

In this section, modelling and control of university offices temperature was investigated in simulation. The control methods were chosen to be potentially applicable with the prototype discussed above.

## 4.1 Model Description

The modelled room (Figure 4) corresponds to a University office, where one of the prototypes is installed, and is 10m long, with a north/south orientation. To represent the thermal behaviour of this room a dynamic model is developed as shown in Equation (1):

$$\frac{\partial T}{\partial t} = \sum_{i=\{x,y,z\}} \left\{ a_i(x,y,z) \frac{\partial^2 T}{\partial i^2} + \frac{h_i(x,y,z)}{\rho_i(x,y,z) Cp_i} \frac{\partial T}{\partial i} \right\} + \sum_i \frac{a_{pi}(x,y,z)}{\lambda_{pi}(x,y,z)} P_i(x,y,z) \quad (1)$$

Where: $(\lambda/\rho C_p)$ is the diffusivity coefficient ($m^2/s$), $\lambda$ is the conduction coefficient (W/m.K), $\rho$ is the density ($kg/m^3$), $C_p$ the calorific capacity (J/kg.K), h stands for the convection coefficient ($W/m^2.K$) and $P_i$ is power density of the $i^{th}$ heat source ($W/m^3$). In order to fine down equation (1), the room is supposed to be constituted by a homogenous and isotropic material, and y- and z-axes are assumed to have infinite lengths. Thus, equation (1) becomes:

$$\frac{\partial T}{\partial t} = a_x \frac{\partial^2 T}{\partial x^2} + \frac{h}{\rho Cp} \frac{\partial T}{\partial x} + \sum_i \frac{a_{xi}}{\rho_i Cp_i} P_i \quad (2)$$

The Crank-Nicholson discrimination method was preferred due to the increased simulation stability and the reduced truncation error (Nougier, 1993). One-dimension heat propagation was considered to promote the preferential direction. External conditions influence the front temperature of the walls by convection, as can be seen in Equation (3):

$$\frac{\partial T}{\partial x} = \frac{h \Delta T}{\rho Cp} \quad (3)$$

Model parameters used were (Sacadura, 1993): air diffusivity coefficient: $2.22.10^{-5}$ $m^2/s$, concrete diffusivity coefficient: $4.2.10^{-5}$ $m^2/s$, air conductivity coefficient: 0.03 W/m.K, indoor and outdoor convection coefficients 10 and 30 $W/m^2.K$, respectively, air density: 1.177 $kg/m^3$ and air specific heat: 1.006 kJ/kg.K. Open loop simulations were performed using real external temperature measurements and constant and equal powers (396W). Figure 5 presents the simulation results, using real external temperature data. Note that walls play the role of linear filters, which explains the stability of the indoor temperature profile. Energy consumptions of the warmers were constant and equal to 792 $Wh/m^2$.

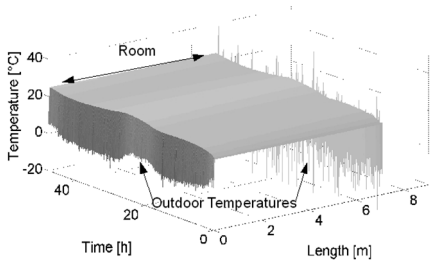



Figure 5: Room Temperature Profile with Open-Loop Control.

## 4.2 Model Predictive Control

Model Predictive Control (García, 1989) is a process control method that uses: (i) a dynamic model of the process, (ii) past control history and (iii) cost optimization over a prediction horizon $H_p$, as shown in Figure 6.

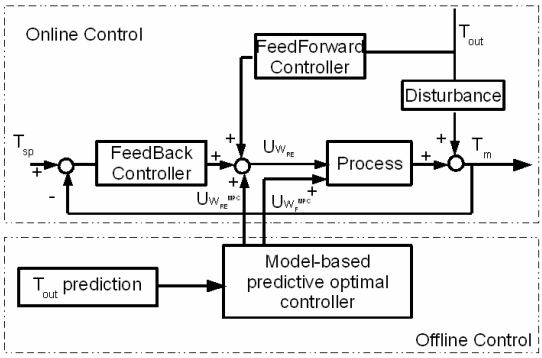



Figure 6: Mixed FB/FF and Predictive Control Scheme.

Such an approach was already tested in this context, but using static modelling, (Kalagasidis, 2006). The corresponding optimization problem is formulated as follows:

$$\min_{U_{WF},U_{WRE}^{MPC}}\left(\sum_{k=1}^{Hp}\left(U_{WF}(k)\right)^2\right)$$

s.t. : Equations (2) - (3)

$$\begin{aligned} U_{W_F}^{\min} &\le U_{W_F}(t) \le U_{W_F}^{\max} \\ U_{W_{RE}}^{\min} &\le U_{W_{RE}}(t) \le U_{W_{RE}}^{\max} \\ \left|T_m(H_p) - T_{sp}(H_p)\right| &= 0 \\ \left|T_m(H_c) - T_{sp}(H_c)\right| &= 0 \end{aligned} \quad (4)$$

Where $U_{W_F}$ and $U_{W_{RE}}^{MPC}$ are the power applied to $W_F$ and an extra-power applied to $W_{RE}$. The idea herein is to use $W_{RE}$ upon saturation before using $W_F$. Hence, $\forall t$, $U_{W_{RE}}(t) = U_{W_{RE}}^{PI+FF}(t) + U_{W_{RE}}^{MPC}(t)$, where $U_{W_{RE}}^{PI+FF}$ is the contribution to the power applied to $W_{RE}$ computed by the on-line controller. The advantage of this formulation is that, while $U_{W_{RE}}(t) \le U_{W_{RE}}^{\max}, U_{W_F} = 0$ for optimality. It is imposed that the room temperature reaches its setpoint at $H_c$ and $H_p$ while minimizing $U_{W_F}$ (Equation 4). $W_{RE}$ is controlled through online Feedback/Feedforward Control, while $W_F$ power increments are computed by MPC, using $H_p = 2\text{h}$ and $H_c = 1\text{h}30$. The optimization problem uses biased external temperatures predictions by means of a 1°C oscillating prediction error.

Figures 7 and 8 show the temperatures time profiles and the powers applied to the warmers, respectively, and Table 1 summarizes energy consumption for the investigated scenarios. Feedback/Feedforward (FB/FF) of the two warmers, for which priority is given to $W_{RE}$, was also investigated for comparison purposes. It seems that most of the reduction is due to the use of time-varying setpoint (see FB/FF results), while setpoint tracking is efficiently achieved. However, this table shows that MPC allows a 7% additional fossil energy consumption reduction when compared to FB/FF.

Table 1: Performance Indicator Values for the Different Control Strategies.

| | Open-loop | FB/FF | FB/FF+MPC |
|---|---|---|---|
| $W_{RE}$ | 792 | 1223.3 | 1227.6 |
| $W_F$ | 792 | 100.9 | 93.5 |

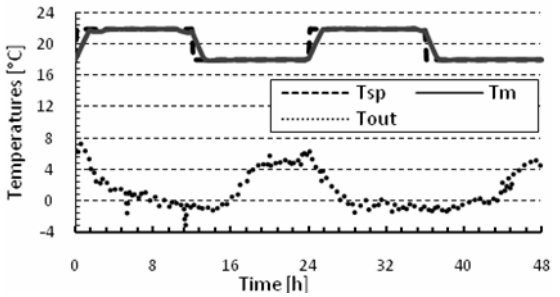



Figure 7: Room, Setpoint and Outdoor Temperatures for FB/FF+MPC Control.

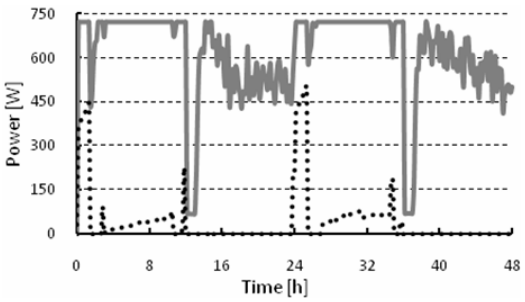



Figure 8: Power Profiles Applied to the Warmers for FB/FF+MPC Control ($W_{RE}$: solid line; $W_F$: dotted).

# 5 CONCLUSIONS

This article presents the results of a study dealing with the improvement of energetic performance of renewable energy buildings. A performance indicator (kWh/m$^2$/yr) was chosen that allows comparisons between buildings of different areas and localizations. A processor-based prototype was developed, to perform on-line acquisition, monitoring and control of heat consumption in renewable energy buildings. The potential for the fossil energy consumption reduction is illustrated by the simulation of temperature control of University's offices. Mixed online and model-based predictive control using both external temperature predictions and real measurements with time-varying temperature setpoint leads to a very large fossil energy consumption reduction.

Future work will include the improvement of the dynamic model, so as to test the developed control algorithms on larger and more complex dynamic systems. Furthermore, in-situ application of the prototype has already begun in our partner's headquarters. It is planned to include control algorithms in addition to real-time data-acquisition and performance indicator monitoring.

# ACKNOWLEDGEMENTS

This work is supported by a fund from the FCE (Funds for the Competitiveness of the Enterprises, DERBI cluster). The authors would like to thank Apex BP Solar, CSTB and Pyrescom for our collaboration and their involvement in this project.

# REFERENCES

ADEME (Agence Départementale de l'Environnement et de la Maîtrise de l'Energie), 2007. *Les chiffres clés du bâtiment en 2006*, Publications de l'ADEME.

Boyde, G., Dutrow, E., Tunnessen, W., 2007. *The evolution of the ENERGY STAR® energy performance indicator for benchmarking industrial plant manufacturing energy use*, J Clean Prod (2007), doi: 10.1016/j.jclepro., 2007.02.024.

Duffaure-Gallais, I., 2006. *La nouvelle réglementation thermique arrive*. Le Moniteur, pages: 16-17.

García, C.E., Prett, D.M., Morari M., 1989: *Model predictive control: Theory and practice - A survey, Automatica*, Vol. 25, n°3, Pages 335-348.

Gungor, M.B., Kyriacou, P.A., Jones, D.P., 1997. *A new micro-controller based wear-time monitor for use with removable orthodontic appliances*. In Proceedings of the 19th Engineering in Medicine and Biology society, Vol. 6, pages: 2419 – 2421.

Hensen, J.L.M., 1991. *On the thermal interaction of building structure and heating and ventilation system*, Ph.D. Thesis Technische Univ., Eindhoven (Netherlands).

Journal Officiel, 2006. *Arrêté du 27 juillet 2006 relatif au contenu et aux conditions d'attribution du label HPE*, République française.

Kalagasidis, A.S., Taesler, R., Andersson C., Nord, M., 2006: *Upgraded weather forecast control of building heating systems*. In Proceedings of the 3rd International Building Physics Conference, Concordia University, Montreal, Canada, pages: 951-958.

Kalogirou, S., Bojic, M., 2000. *Artificial neural networks for prediction of the energy consumption of a passive solar building*, Energy, Vol. 25, n°5, Pages 479-491.

Kalogirou, S., Lalot, S., Florides, G., Desmet, B., 2007. *Development of neural network-based fault diagnostic system for solar thermal applications*, Sol. Energy, doi: 10.1016/j.solener.2007.06.010.

Leong, S.S., Vun, C.H., 1998, *Design and implementation of an authentication protocol for home automation systems*, IEEE Transactions on Consumer Electronics, Vol. 44, n°3, pages: 911-921.

Lygouras, J.N., Botsaris, P.N., Vourvoulakis, J., Kodogiannis, V., 2006. *Fuzzy logic controller implementation for a solar air-conditioning system*, Ap. Energy, doi: 10.1016/j.apenergy.2006.10.002.

Maïzi, N., Assoumou, E., 2007. *Modélisation prospective et spécificités de la politique énergétique française,* Journal sur l'enseignement des sciences et technologies de l'information et des systèmes, vol. 6, doi :10.1051/J3EA :2007002.

Mendoça, P., Bragança, L., 2003. *Energy optimization through thermal zoning the outer skin*, In Proceedings of the Healthy Buildings 7th International Conference, School of Design and Environment, National Univ. of Singapore.

Mendoça, P., Bragança, L., 2007. *Sustainable housing with mixedweight strategy - A case of study*. Building and Environment, doi: 10.1016/j.buildenv., 2006.08.025.

Nougier, J.P., 1993. *Méthodes de calculs numériques*, Masson.

Sacadura, J.F., 1993. *Initiation aux transferts thermiques*, Edition Technique et Documentation.

Segars, S., 1998. *The ARM9 family-high performance microprocessors for embedded applications*, In Proceedings of the International Conference on Computer Design: VLSI in Computers and Processors, pages: 230-235.

Sesolis, B, 2006. *Amélioration de la performance énergétique : Une nouvelle réglementation pour tous les bâtiments neufs, la RT 2005*, JMG.

Xingwu, C., Xinhua, J., Lei, W., 2006. *Development on ARM9 System-on-chip Embedded Sensor Node for Urban Intelligent Transportation System*. In Proceedings of the International Symposium on Industrial Electronics, IEEE, Vol. 4, Pages: 3270-3275.

# ASYMPTOTIC THEORY OF THE REACHABLE SETS TO LINEAR PERIODIC IMPULSIVE CONTROL SYSTEMS

E. V. Goncharova
*Institute for System Dynamics and Control Theory, Siberian Branch of Russian Academy of Sciences*
*134 Lermontov Street, Irkutsk, Russia*
*goncha@icc.ru*

A. I. Ovseevich
*Institute for Problems in Mechanics, Russian Academy of Sciences*
*101 Vernadsky Avenue, Moscow, Russia*
*ovseev@ipmnet.ru*

Keywords: Linear periodic dynamic systems, impulsive control, reachable sets, shapes of convex bodies.

Abstract: We study linear periodic control systems with a bounded total impulse of control. The main result is an asymptotic formula for the reachable set, which, at the same time, reveals the structure of the attractor — the set of all limit shapes of the reachable sets. The attractor is shown to be parameterized by a (finite-dimensional) toric fibre bundle over a circle. The fibre of the bundle can be described via the Floquet multipliers (monodromy matrix) of the linear system. Moreover, the limit dynamics of shapes of reachable sets can be parametrized by an explicit curve on the toric bundle.

## 1 INTRODUCTION

One of the fundamental notions of control theory is that of reachable sets which provide a visible bound for control capabilities. In general, these sets have a complicated shape and dynamics. There is, however, a kind of problems where the behavior of reachable sets is well-understood.

Namely, it turns out that the reachable sets of linear control systems have simple limit properties as time evolves to infinity provided that a suitable time-dependent matrix scaling is applied. This kind of results was found for the first time in (Ovseevich, 1991), where time-invariant linear control systems with geometric bounds on control were studied. It was shown that in this setup there is a single limit shape of the reachable sets, shape being the set regarded up to an arbitrary nondegenerate linear transform.

At present the scope of this phenomena is not yet clear cut. It is very likely that there is a natural extension of these results to general time-dependent linear systems. Moreover, a similar phenomena was discovered for some nonlinear stochastic dynamic systems (Dolgopyat et al., 2004).

The purpose of the study is to develop the asymptotic theory of the reachable sets to linear impulsive control systems. A motivation to address impulsive control systems is also due to the perceived relevance of the impulsive control theory for hybrid systems whose state evolution is dictated by the interaction of conventional time-driven dynamics and event-driven dynamics (see, e.g., (Aubin, 2000; Branicky et al., 1998; Miller and Rubinovich, 2003)).

In this paper, we study the periodic linear control systems with a bounded total impulse of control. The main result is an asymptotic formula for the reachable sets (see (3)), that, in particular, reveals the structure of attractor — the set of all limit shapes of the reachable sets.

It would be extremely interesting to understand the limit behavior of reachable sets for a general linear system. Unfortunately, the nature of the present methods is computational and it looks like new ideas are needed in order to grasp the limit dynamics of the reachable sets.

## 2 PROBLEM STATEMENT

Consider a linear control system on the time interval $[0,T]$

$$\dot{x}(t) = A(t)x(t) + B(t)u(t),\ x(0) = 0, \quad (1)$$

under the constraint on the total impulse of control $u$:

$$\int_0^T \langle f(t), u(t)dt \rangle \leq 1, \qquad (2)$$

where $x(t) \in \mathbb{V} = \mathbb{R}^n$, $u(t) \in \mathbb{W} = \mathbb{R}^m$, $A(t)$, $B(t)$ are matrices of appropriate dimensions, $U(t)$ is a given central symmetric convex body in $\mathbb{W}$, $f$ is an arbitrary continuous function such that $f(t) \in U^\circ(t)$, and $U^\circ(t)$ is the polar of set $U(t)$.

Assume that the Kalman type condition of complete controllability holds, namely, for any vector $u \in \mathbb{W}$ and a time moment $T$, function $\Phi(T,t)B(t)u$ does not vanish identically in any interval of time $t$. Under these assumptions the reachable sets $\mathcal{D}(T)$ to system (1), (2) are central symmetric convex bodies.

The problem addressed is to study the limit behavior of the reachable sets $\mathcal{D}(T)$ as $T \to \infty$. The reachable sets are regarded as elements of the metric space $\mathbb{B}$ of central symmetric convex bodies with the Banach-Mazur distance $\rho$:

$$\rho(\Omega_1, \Omega_2) = \log(t(\Omega_1, \Omega_2)t(\Omega_2, \Omega_1)),$$
$$\text{where } t(\Omega_1, \Omega_2) = \inf\{t \geq 1 : t\Omega_1 \supset \Omega_2\}.$$

The general linear group $GL(\mathbb{V})$ naturally acts on the space $\mathbb{B}$ by isometries. The factorspace $\mathbb{S}$ is called the space of shapes of central symmetric convex bodies, where the shape $\mathrm{Sh}\,\Omega \in \mathbb{S}$ of a convex body $\Omega \in \mathbb{B}$ is the orbit $\mathrm{Sh}\,\Omega = \{C\Omega : \det C \neq 0\}$ of the point $\Omega$ with respect to the action of $GL(\mathbb{V})$. The Banach-Mazur factormetric makes $\mathbb{S}$ into a compact metric space. The convergence of the reachable sets $\mathcal{D}(T)$ and their shapes is understood in the sense of the Banach-Mazur metric. For two asymptotically equal functions with values in the space of convex bodies or the space of their shapes, the following notations are used: $\Omega_1(T) \sim \Omega_2(T)$, if $\rho(\Omega_1(T), \Omega_2(T)) \to 0$ as $T \to \infty$, and similarly $\mathrm{Sh}\,\Omega_1(T) \sim \mathrm{Sh}\,\Omega_2(T)$, if $\rho(\mathrm{Sh}\,\Omega_1(T), \mathrm{Sh}\,\Omega_2(T)) \to 0$ as $T \to \infty$. The convergence of convex bodies may be also understood in the sense of convergence of their support functions. Remind that the support function of a convex compact set is given by formula: $H_\Omega(\xi) = \sup_{x \in \Omega} \langle x, \xi \rangle$, where $\xi \in \mathbb{V}^*$, and uniquely defines the set $\Omega$. The equivalence of the two definitions of convergence of convex bodies — in the terms of convergence of their support functions and in the sense of the Banach-Mazur metric — is established by the following lemma (Figurina and Ovseevich, 1999):

**Lemma 1.** *A sequence $\Omega_i \in \mathbb{B}$ converges to $\Omega \in \mathbb{B}$ in the sense of the Banach-Mazur metric if and only if the corresponding sequence of the support functions $H_i(\xi) = H_{\Omega_i}(\xi)$ converges to the support function $H_\Omega(\xi)$ pointwise and is uniformly bounded on the unit sphere in the dual space $\mathbb{V}^*$.*

We address the periodic case, when the constituents $A$, $B$, and $U$ of control system (1), (2) are supposed to be continuous and periodic in $t$. To fix ideas, the period is assumed to be 1. It would be interesting to understand the limit behavior of reachable sets for a general linear system. This problem, however, seems rather difficult, since already the time-invariant case is nontrivial, and, say, for quasi-periodic systems it is not clear how to prove the corresponding natural conjectures.

# 3 ASYMPTOTIC BEHAVIOR OF SHAPES OF THE REACHABLE SETS

We study the limit behavior as $T \to +\infty$ of the curve $T \mapsto \mathrm{Sh}\,\mathcal{D}(T)$ under different assumptions on the spectrum of the monodromy matrix. At the heart of the considerations below there is an explicit formula for the support function of the reachable set:

**Lemma 2.** *The support function of the reachable set $\mathcal{D}(T)$ to system (1), (2) is given by*

$$H_{\mathcal{D}(T)}(\xi) = \sup_{t \in [0,T]} H_{U(t)}\left(B(t)^* \Phi(T,t)^* \xi\right), \qquad (3)$$

*where $\Phi(t,s)$ is the fundamental matrix of linear system $\dot{x} = A(t)x$.*

**Stable Case.** Let the system (1) be asymptotically stable, i.e.

$$\Phi(T,t) = o(1) \text{ as } T - t \to +\infty,$$

and $o(1)$ is uniformly small. It is easy to establish the stability criterion: system (1) is asymptotically stable iff the spectrum of the monodromy matrix $M = \Phi(1,0)$ is contained in the open unit disk of the complex plane.

Let us show that the curve $T \mapsto \mathcal{D}(T)$ is asymptotically periodic as $T \to \infty$. In other words, there exists such a continuous periodic curve $f : \mathbb{R}/\mathbb{Z} \to \mathbb{B}$ that $\mathcal{D}(T) \sim f(T)$ as $T \to +\infty$. Informally speaking, the curve $T \mapsto \mathcal{D}(T)$ is reeled on a limit cycle.

The curve $f$ can be given by an explicit formula. Define function

$$\mathcal{F}(T) = \mathcal{F}(T, \xi) = \sup_{t \in (-\infty, T]} H_U\left(B^* \Phi(T,t)^* \xi\right), \qquad (4)$$

where the argument $t$ of periodic functions $B$ and $U$ is omitted. Due to the stability condition, $\mathcal{F}(T)$ is a continuous periodic function of $T$. The periodicity of $\mathcal{F}$ follows from the equality

$$\Phi(T+1, t+1) = \Phi(T,t)$$

for the fundamental matrix of a 1-periodic system. Furthermore, for each $T$, function $\xi \mapsto \mathcal{F}(T,\xi)$ is homogeneous and convex, and therefore it is the support function of a body $f(T)$. Thus, the curve $f$ is defined. On the other hand, from (3), (4) it follows that

$$H_{f(T)}(\xi) = H_{\mathcal{D}(T)}(\xi) \text{ for large enough } T, \quad (5)$$

since $\Phi(T,t) = O\left(e^{-\beta(T-t)}\right)$, where $\beta > 0$ owing to the assumed stability property. The asymptotic equality $\mathcal{D}(T) \sim f(T)$ follows from (5).

**Unstable Case.** Assume that the system (1) is strictly unstable, i.e.

$$\Phi(T,s) = o(1) \text{ as } T - s \to -\infty,$$

where $o(1)$ is uniformly small. System (1) is strictly unstable iff the spectrum of the monodromy matrix $M = \Phi(1,0)$ is contained in the complement of the closed unit disk of the complex plane.

Define the matrix multiplier $C(T) = \Phi(0,T)$ and consider the set

$$\widetilde{\mathcal{D}}(T) \stackrel{\text{def}}{=} C(T)\mathcal{D}(T).$$

It is easy to see that

$$H_{\widetilde{\mathcal{D}}(T)}(\xi) = \sup_{t\in[0,T]} H_U\left(B^*\Phi(0,t)^*\xi\right),$$

and from the instability criterion it follows that $\Phi(0,t)$ decreases exponentially fast as $t \to \infty$. Therefore, the values $\sup\limits_{t\in[0,T]} H_U\left(B^*\Phi(0,t)^*\xi\right)$ converge as $T \to \infty$, and the convergence of bodies $\widetilde{\mathcal{D}}(T) \to \mathcal{D}_\infty$, takes place, where

$$H_{\mathcal{D}_\infty}(\xi) = \sup_{t\in[0,\infty]} H_U\left(B^*\Phi(0,t)^*\xi\right).$$

Thus, we have the asymptotical equality:

$$\mathcal{D}(T) \sim \Phi(T,0)\mathcal{D}_\infty.$$

In the view of the behavior of shapes of the reachable sets, there is an essential difference between stable and unstable cases. In the unstable case, shapes $\operatorname{Sh}\mathcal{D}(T)$ converge as $T \to \infty$, while, for stable system, the curve $T \to \operatorname{Sh}\mathcal{D}(T)$ is reeled on a limit cycle.

Note that the above considerations admit an extension to almost periodic systems.

Yet another choice $C(T) = \Phi(\{T\},T)$, where $\{T\}$ is the fractional part of $T$, of a normalizing matrix factor seems also reasonable. It is easy to see at that rate, that the limit normalized body $\mathcal{D}_\infty = \mathcal{D}_\infty(T)$ depends on time periodically. This choice of the matrix factor fits better the general case, when stable, unstable and neutral components are present.

**Neutral Case.** Assume that system (1) is neutral, meaning that the spectrum of the monodromy matrix $M = \Phi(1,0)$ rests on the unit circle. Consider the Jordan decomposition

$$M = U\mathfrak{D} = e^{N+D},$$

where $\mathfrak{D}$ is a diagonalizable matrix of the same spectrum that the matrix $M$ has, $D$ is such a diagonalizable real matrix that $\mathfrak{D} = e^D$, $N$ is a nilpotent matrix, and $ND = DN$. As is well known, there exists a matrix $F = F(N,T)$ with the following properties:

$$FNF^{-1} = T^{-1}N,\ FD = DF,$$
$$\text{and } F_\infty = F_\infty(N) = \lim_{T\to+\infty} F(N,T) \text{ is defined.}$$

Put

$$N(T) = \Phi(T,0)N\Phi(0,T),\ D(T) = \Phi(T,0)D\Phi(0,T).$$

It is easy to see that $N(T)$ and $D(T)$ are periodic functions of $T$, since matrices $N$ and $D$ commute with $M = \Phi(1,0)$. Matrix function

$$F_\infty(N(T)) = \Phi(T,0)F_\infty(N)\Phi(0,T)$$

is also continuous and periodic. Function $\phi$ given by formula

$$\phi(T,t) = e^{(t-T)[N(T)+D(T)]}\Phi(T,t), \quad (6)$$

is periodic in $T$ and $t$. Define the matrix factor

$$C(T) = F(N(T),T)e^{T[N(T)+D(T)]}$$

and consider the normalized set

$$\widetilde{\mathcal{D}}(T) \stackrel{\text{def}}{=} C(T)\mathcal{D}(T).$$

It is easy to see that

$$\begin{aligned} &H_{\widetilde{\mathcal{D}}(T)}(\xi) = \sup H\left(g_T^1(t)\right) = \sup H\left(g_T^2(t)\right) \\ &= \sup H\left(g_T(t)\right) + o(1), \end{aligned} \quad (7)$$

where sup is taken over $t \in [0,T]$, $H = H_U$ and

$$g_T^1(t) = B^*\phi(T,t)^*e^{t(N^*(T)+D^*(T))}F(N(T),T)^*\xi,$$
$$g_T^2(t) = B^*\phi(T,t)^*F(N(T),T)^*e^{\frac{t}{T}N^*(T)+tD^*(T)}\xi,$$
$$g_T(t) = B^*\phi(T,t)^*F_\infty(N(T))^*e^{\frac{t}{T}N^*(T)+tD^*(T)}\xi.$$

The last term $o(1)$ in (7) appears on account of the difference between $F(T)$ and $F_\infty(N(T))$. Since other matrices involved in $g_T^2(t)$ are uniformly bounded, the difference of the arguments comes to $o(1)$.

Consider the following function of two arguments $T$ and $L$:

$$I(L,T) = \sup_{t\in[0,L]} f_T(t,\mathbf{t},\tau) =$$
$$= \sup_{t\in[0,L]} H_U\left(B^*\phi(T,t)^*F_\infty(N(T))^*e^{\frac{t}{L}N^*(T)+tD^*(T)}\xi\right).$$

Function $f_T(t,\mathbf{t},\tau)$, where $\mathbf{t}=e^{tD(T)}$, $\tau=t/L$, is periodic in $t$ and depends on the parameter $T$ periodically as well. By using the Hermann Weyl averaging method (Weyl, 1938; Weyl, 1939; Arnold, 1989), like in (Goncharova and Ovseevich, 2007), we obtain the asymptotic representation

$$I(L,T)=\sup_{\mathcal{T}\times J} f_T(t,\mathbf{t},\tau)+o(1) \qquad (8)$$

as $L\to\infty$, where $o(1)$ is small uniformly in $T$. In formula (8), the interval $J=[0,1]$ and a torus $\mathcal{T}=\mathcal{T}(T)$ are involved. The torus is the closure of the one-parameter subgroup $\{(e^{2\pi it},e^{tD(T)})\}$ in the group $S^1\times \mathrm{GL}(\mathbb{V})$. Notice that the torus

$$\mathcal{T}(T)=\Phi(T,0)\mathcal{T}(0)\Phi(0,T)$$

depends on $T$ continuously and periodically. The torus can be naturally represented as a fibre bundle over the circle, at that the fibre over $e^{2\pi iT}\in S^1$ is the closure of the cyclic group generated by matrix $e^{D(T)}\in\mathrm{GL}(\mathbb{V})$.

It is clear that

$$I(T)=\sup_{\mathcal{T}\times J} f_T(t,\mathbf{t},\tau)$$

is a periodic function of parameter $T$. Hence, for large $L$,

$$I(L,T)=I(T)+o(1)$$

is a periodic function of $T$ up to $o(1)$. In particular, this is true for $L=T$ and large $T$. From this we conclude that the curve $T\mapsto \mathrm{Sh}\,\mathcal{D}(T)$ is periodic up to $o(1)$, i.e. it is reeled on a limit cycle.

**Stable-neutral Case.** Suppose that system (1) is stable-neutral. This means that fundamental matrix admits a polynomial estimate:

$$|\Phi(T,t)|=O(1+|T-t|^n)\ \text{ as } T-t\to+\infty.$$

It is not difficult to obtain the criterion of stable-neutrality: system (1) is stable-neutral iff the spectrum of the monodromy matrix $M=\Phi(1,0)$ is contained in the closed unit disk of the complex plane.

Consider the canonical decomposition of the monodromy matrix $M=\Phi(1,0)$

$$M=M_0\oplus M_-$$

into the stable and neutral components (in accordance with the relations $|\lambda|<1$, $|\lambda|=1$ for eigenvalues), and the corresponding decomposition of phase space:

$$\mathbb{V}=\mathbb{V}_0\oplus\mathbb{V}_-.$$

For an arbitrary time moment $T$, the monodromy matrix

$$M(T)=\Phi(T+1,T)=\Phi(T,0)M\Phi(0,T)$$

depends on $T$ periodically. The corresponding decomposition

$$\mathbb{V}(T)=\mathbb{V}_0(T)\oplus\mathbb{V}_-(T).$$

is also periodic in $T$.

The scaling matric factor $C(T)$ can be taken in the block-diagonal form

$$C(T)=C_0(T)\oplus C_-(T),$$

where $C_i(T):\mathbb{V}_i(T)\to\mathbb{V}_i(T)$ are given by formulas

$$C_0(T)=F(N(T),T),\ C_-(T)=I.$$

The support function $H_{\widetilde{\mathcal{D}}(T)}(\xi)$ of the normalized body $\widetilde{\mathcal{D}}(T)\stackrel{\mathrm{def}}{=}C(T)\mathcal{D}(T)$ is as follows

$$H_{\widetilde{\mathcal{D}}(T)}(\xi)=\sup_{t\in[0,T]} f_T(t)=\sup_{t\in[0,T]} H_U(g_T(t)), \qquad (9)$$

$$g_T(t)=B^*\Phi(T,t)^*\xi_-+$$

$$B^*\phi(T,t)^*e^{(T-t)(N^*(T)+D^*(T))}F(N(T),T)^*\xi_0,$$

$\xi_i=\xi_i(T)\in\mathbb{V}_i(T)^*$, $i\in\{-,0\}$ are components of the canonical decomposition of a vector $\xi\in\mathbb{V}^*$, and the periodic in both arguments function $\phi(T,t)$ is defined in (6). Notice that in the generic case, matrix $F$ is the identity one, and therefore, the formula for the support function is just as simple as (3):

$$H_{\widetilde{\mathcal{D}}(T)}(\xi)=\sup_{t\in[0,T]} H_U\left(B^*\Phi(T,t)^*\xi\right).$$

To study $H_{\widetilde{\mathcal{D}}(T)}(\xi)$ as $T\to\infty$ let us apply the decomposition method like in the autonomous case (Goncharova and Ovseevich, 2007). Owing to the basic commutativity relations for matrices $F$, $N$, and $D$, we have

$$e^{(T-t)(N^*(T)+D^*(T))}F(N(T),T)^*\xi_0=$$
$$=F(N(T),T)^*e^{\frac{T-t}{T}N^*(T)+(T-t)D^*(T)}\xi_0,$$

and, thus, we have the uniform asymptotic equality

$$\Phi(T,t)^*F(N(T),T)^*\xi_0=$$
$$\phi(T,t)^*F_\infty(N(T))^*e^{\frac{T-t}{T}N^*(T)+(T-t)D^*(T)}\xi_0+o(1)$$

as $T\to\infty$. Like in (Goncharova and Ovseevich, 2007), we divide the time interval $\mathit{I}=[0,T]$ into the two subintervals

$$\mathit{I}=\mathit{I}_0\cup\mathit{I}_-=[0,(1-\varepsilon)T]\cup[(1-\varepsilon)T,T],$$

where $\varepsilon=\varepsilon(T)$ is such that $\varepsilon(T)=o(1)$, while $\varepsilon(T)T\to\infty$ as $T\to\infty$. By adopting arguments from (Goncharova and Ovseevich, 2007), we obtain the asymptotic equality

$$H_{\widetilde{\mathcal{D}}(T)}(\xi)=\max\{H_{-0}(T,\xi),H_0(T,\xi)\}+o(1), \qquad (10)$$

where in accordance with notations (9)

$$H_{-0} = \sup_{t\in I_-} f_T(t) + o(1) = \sup_{t\in(-\infty,T]} H_U(g_{-0T}(t)),$$

where $g_{-0T}(t)$ stands for

$$B^*\Phi(T,t)^*\xi_- + B^*\phi(T,t)^*F_\infty(N(T))^*e^{(T-t)D^*(T)}\xi_0,$$

while

$$H_0 = \sup_{t\in I_0} f_T(t) + o(1) = \sup_{t\in\mathbf{R},\tau\in[0,1]} H_U(g_{0T}),$$

where $g_{0T}(t,\tau)$ stands for

$$B^*\phi(T,t)^*F_\infty(N(T))^*e^{\tau N^*(T)+(T-t)D^*(T)}\xi_0.$$

Functions $H_{-0}(T,\xi)$ and $H_0(T,\xi)$ are periodic in $T$, and convex, homogeneous in $\xi$. From this it follows that $H_i(T,\xi)$, $i=-0,0$, are the support functions of some convex compact sets $\Omega_{-0}(T)\subset\mathbb{V}(T)$ and $\Omega_0(T)\subset\mathbb{V}_0(T)$, which periodically depend on $T$. Furthermore, the convex compact

$$\Omega(T) = \Omega_{-0}(T) * \Omega_0(T) \subset \mathbb{V}(T),$$

and $\Omega_0(T)$) is a body with the support function

$$\max\{H_{-0}(T,\xi), H_0(T,\xi)\},$$

and also periodically depends on $T$. Here, we use the join notation:

$$\Omega = \Omega' * \Omega'',$$

meaning that $\Omega$ is the convex hull of the union $\Omega'\bigcup\Omega''$, or what is the same $H_\Omega = \max\{H_{\Omega'}, H_{\Omega''}\}$. Thus, the curve $T\mapsto\widetilde{\mathcal{D}}(T)$ is reeled on the limit cycle $T\mapsto\Omega(T)$.

**Unstable-neutral Case.** Similarly to stable-neutral case, fundamental matrix admits a polynomial estimate

$$|\Phi(T,t)| = O(1+|T-t|^n) \text{ as } T-t\to-\infty.$$

System (1) is unstable-neutral iff the spectrum of the monodromy matrix $M=\Phi(1,0)$ is contained in the closed complement of the unit disk of the complex plane. In this case, the normalizing matrix factor $C(T)$ can be taken in the block-diagonal form:

$$C(T) = C_0(T)\oplus C_+(T),$$

where $C_i(T):\mathbb{V}_i(T)\to\mathbb{V}_i(T)$ are defined by formulas

$$C_0(T) = F(N(T),T),\ C_+(T) = \Phi(\{T\},T).$$

We note in passing that if we put $C_+(T)=\Phi(0,T)$, then the block-diagonal structure of the normalizing matrix would be lost, since, in general, $\Phi(0,T)$ do not map $\mathbb{V}_+(T)$ into itself.

The support function $H_{\widetilde{\mathcal{D}}(T)}(\xi)$ of the normalized body $\widetilde{\mathcal{D}}(T)\stackrel{\text{def}}{=}C(T)\mathcal{D}(T)$ is similar to (9):

$$\sup_{t\in[0,T]} H_U(B^*\Phi(\{T\},t)^*\xi_+ + B^*\phi(T,t)^*\times$$
$$\times e^{(T-t)(N^*(T)+D^*(T))}F(N(T),T)^*\xi_0),$$

and like in (10) we have the asymptotics

$$H_{\widetilde{\mathcal{D}}(T)}(\xi) = \max\{H_{+0}(T,\xi), H_0(T,\xi)\} + o(1),$$

where $H_{+0}(T,\xi) = \sup_{t\in[0,\infty)} H_U(g_{+0T}(t))$, and

$$g_{+0T}(t) = B^*\Phi(\{T\},t)^*\xi_+ +$$
$$+B^*\phi(T,t)^*F_\infty(N(T))^*e^{N(T)^*}e^{(T-t)D^*(T)}\xi_0.$$

The essential difference, in contrast to stable-neutral situation, consists of that, on this occasion, function $H_{+0}(T,\xi)$ is not periodic in $T$, however, it becomes periodic by the transformation of variable $\xi_0\mapsto e^{TD^*(T)}\xi_0$. Geometrically, this means that the asymptotic equality

$$\widetilde{\mathcal{D}}(T) \sim p(\mathbf{t}_T)\Omega_{+0}(T) * \Omega_0(T) \text{ as } T\to\infty,$$

holds, where $\mathbf{t}_T = (e^{2\pi iT}, e^{TD(T)})$ is an element of the torus $\mathcal{T}(T)$, matrix $p(\mathbf{t}_T) = e^{TD(T)}$ is the second component of $\mathbf{t}_T$, $\Omega_\alpha(T)$ are periodically depending on time convex compacts in spaces $V_\alpha(T)$, $\alpha\in\{+0,0\}$.

Thus, the limit behavior of the normalized reachable sets $\widetilde{\mathcal{D}}(T)$ is the same as the behavior of the curve $T\mapsto\mathbf{t}_T$. The closure of the curve might have an arbitrary large dimension so that the curve is reeled on a multidimensional manifold. Still, the shapes of the reachable sets $\mathrm{Sh}\,\mathcal{D}(T)$ have a simpler behavior, since $\mathrm{Sh}\,\mathcal{D}(T)\sim\mathrm{Sh}\,(\Omega_{+0}(T)*\Omega_0(T))$, and, therefore, the curve $T\mapsto\mathrm{Sh}\,\mathcal{D}(T)$ is reeled on a limit cycle of dimension not greater than 1.

**General Case.** The general result can be obtained by using the decomposition method (see (Goncharova and Ovseevich, 2007)) and the considered above cases. Consider the canonical decomposition of the monodromy matrix $M=\Phi(1,0)$

$$M = M_+\oplus M_0\oplus M_-$$

into the unstable, neutral, and stable components ( in accordance with the relations $|\lambda|<1$, $|\lambda|>1$, $|\lambda|=1$ for eigenvalues), and the corresponding decomposition of phase space

$$\mathbb{V} = \mathbb{V}_+\oplus\mathbb{V}_0\oplus\mathbb{V}_-.$$

For an arbitrary time moment $T$, the monodromy matrix

$$M(T) = \Phi(T+1,T) = \Phi(T,0)M\Phi(0,T)$$

depends on $T$ periodically, so does the corresponding decomposition of phase space

$$\mathbb{V} = \mathbb{V}_+(T) \oplus \mathbb{V}_0(T) \oplus \mathbb{V}_-(T).$$

In the general case, the scaling matric factor $C(T)$ can be chosen in the block-diagonal form

$$C(T) = C_+(T) \oplus C_0(T) \oplus C_-(T),$$

where $C_i(T) : \mathbb{V}_i(T) \to \mathbb{V}_i(T)$ are given by formulas

$$C_+(T) = \Phi(\{T\}, T),$$
$$C_0(T) = F(N(T), T),\ C_-(T) = I.$$

The normalized body $\widetilde{\mathcal{D}}(T) = C(T)\mathcal{D}(T)$ has the following asymptotics

$$\widetilde{\mathcal{D}}(T) \sim p(\mathbf{t})\Omega_{+0}(T) * \Omega_0(T) * \Omega_{-0}(T) \qquad (11)$$

as $T \to \infty$, where $\mathbf{t} = (e^{2\pi i T}, e^{TD(T)})$ is an element of the torus $\mathcal{T}(T)$, matrix $p(\mathbf{t}) = e^{TD(T)}$ is the second component of $\mathbf{t}$, $\Omega_\alpha(T)$ are periodically depending on time convex compacts in $V_\alpha(T)$, $\alpha \in \{+0, 0\}$. Asymptotic equality (11) can be naturally interpreted in the terms of attractors in space $\mathbb{S}$ of shapes of convex bodies. Define a fibre bundle over circle $\mathcal{P} \to S^1$ as follows:

$$\mathcal{P} = \{(e^{2\pi i T}, e^{TD(T)} z) \in S^1 \times \mathrm{GL}(\mathbb{V}) : z \in \mathcal{Z}(T)\},$$
$$(e^{2\pi i T}, e^{TD(T)} z) \mapsto e^{2\pi i T},$$

where $\mathcal{Z}(T)$ is the closure in $\mathrm{GL}(\mathbb{V})$ of a cyclic group generated by matrix $e^{D(T)}$. Then, the relation (11) asserts that the totality $\mathcal{A}$ of all the limit shapes of the reachable sets (attractor) is parameterized by the set $\mathcal{P}$: there is a continuous map $\sigma$ from $\mathcal{P}$ onto $\mathcal{A}$. At that, in the limit, the curve $T \mapsto \mathrm{Sh}\,\mathcal{D}(T)$ is parameterized by the curve $T \mapsto \mathbf{t}_T = (e^{2\pi i T}, e^{TD(T)})$ in $\mathcal{P}$ in the sense that

$$\mathrm{Sh}\,\mathcal{D}(T) \sim \sigma(\mathbf{t}_T)$$

as $T \to \infty$.

This asymptotic equality is an incarnation of (11) and the main result of the paper. It says that the limit dynamics of the shapes $\mathrm{Sh}\,\mathcal{D}(T)$ can be described via the "straight winding" $T \mapsto \mathbf{t}_T$ in the toric bundle $\mathcal{P}$.

## 4 CONCLUSIONS

In this paper we determined completely the asymptotic behavior of reachable sets to periodic linear dynamic systems with impulsive control. This is just a single step in the long road directed to understanding the limit behavior of reachable sets for general linear systems. Still, our results for the periodic case suggest a reasonable conjectural description of this behavior. In fact, it is possible to state a precise conjecture pertaining to the quasi-periodic case.

## ACKNOWLEDGEMENTS

The work was partially supported by RFBR (projects 08-01-00156, 08-08-00292).

## REFERENCES

Arnold, V. I. (1989). *Mathematical Methods of Classical Mechanics*. Nauka, Moscow (in Russian).

Aubin, J.-P. (2000). *Impulse differential equations and hybrid systems: A viability approach*. Lecture Notes, University of California at Berkeley.

Branicky, M., Borkar, V. S., and Mitter, S. K. (1998). A unified framework for hybrid control: model and optimal control theory. *IEEE Transactions on Automatic Control*, 43(1):31–45.

Dolgopyat, D., Kaloshin, V., and Koralov, L. (2004). A limit shape theorem for periodic stochastic dispersion. *CPAM*, 57:1127–1158.

Figurina, T. Y. and Ovseevich, A. I. (1999). Asymptotic behavior of attainable sets of linear periodic control systems. *J. Optim. Theory Appl.*, 100(2):349–364.

Goncharova, E. and Ovseevich, A. (2007). Asymptotic behavior of reachable sets of linear impulsive control systems. *J. of Computer & Systems Sciences Intl.*, (1):51–59.

Miller, B. and Rubinovich, E. (2003). *Impulsive Control in Continuous and Discrete-Continuous Systems*. Kluwer Academic Publishers, New York.

Ovseevich, A. I. (1991). Asymptotic behaviour of attainable and superattainable sets. In *Proceedings of the Conference on Modeling, Estimation and Filtering of Systems with Uncertainty, Sopron, Hungary. 1990*, Basel, Switzerland. Birkhaüser.

Weyl, H. (1938). Mean motion. *Amer. J. Math.*, 60:889–896.

Weyl, H. (1939). Mean motion. *Amer. J. Math.*, 61:143–148.

# HETEROGENEOUS IMAGE RETRIEVAL SYSTEM BASED ON FEATURES EXTRACTION AND SVM CLASSIFIER

Rostom Kachouri
*Research unit on Computers, Imaging, Electronics and Systems, ENIS, BP W, 3038 Sfax, Tunisia*
*Informatics, Integrative Biology and Complex Systems, 40 Rue de Pelvoux, 91020 Evry Cedex, France*
*rostom.kachouri@enis.rnu.tn*

Khalifa Djemal, Hichem Maaref
*Informatics, Integrative Biology and Complex Systems, 40 Rue de Pelvoux, 91020 Evry Cedex, France*
*Khalifa.Djemal@iup.univ-evry.fr, maaref@iup.univ-evry.fr*

Dorra Sellami Masmoudi, Nabil Derbel
*Research unit on Computers, Imaging, Electronics and Systems, ENIS, BP W, 3038 Sfax, Tunisia*

Keywords: CBIR, SVM, QUIP-tree, feature extraction, heterogeneous image database.

Abstract: Image databases represent increasingly important volume of information, so it is judicious to develop powerful systems to handle the images, index them, classify them to reach them quickly in these large image databases. In this paper, we propose an heterogeneous image retrieval system based on feature extraction and Support vector machines (SVM) classifier.
For an heterogeneous image database, first of all we extract several feature kinds such as color descriptor, shape descriptor, and texture descriptor. Afterwards we improve the description of these features, by some original methods. Finally we apply an SVM classifier to classify the consequent index database.
For evaluation purposes, using precision/recall curves on an heterogeneous image database, we looked for a comparison of the proposed image retrieval system with an other Content-based image retrieval (CBIR) which is QUadtree-based Index for image retrieval and Pattern search (QUIP-tree). The obtained results show that the proposed system provides good accuracy recognition, and it prove more better than QUIP-tree method.

## 1 INTRODUCTION

Several methods ensuring image recognition were developed. But these techniques are often developed for one kind of image and present difficulties for recognition in an heterogeneous image database.

Different applications domains like medical domain, industrial domain etc, demonstrate a real need for image recognition in large databases. To this end we can distinguish two main types of image databases: the specific database where the images show a natural similarity (the same type of images, the same content presented in a different situation, etc), and heterogeneous databases, which can contain different types and image content. One of the important steps in a recognition system is the image description. Indeed, this step is based on a priori knowledge of the image content on the one hand and on the modeled descriptors for a specific type of image. Methods based on this concept gave satisfaction for specific databases. The relevance of this description strategy becomes almost ineffective when image databases are heterogeneous. It is within this framework, that the system we present is registered. A content image recognition system is typically composed of two main phases, images description and extracted features classification allowing effective recognition.

In fact, in an heterogeneous image database, images are various categories, and we can find a big difference between them. So a unique feature or a unique feature kind, can not be relevant to describe the whole image database. In this paper, we present an heterogeneous image recognition system, to this aim, several kinds of features was used and improved for this purpose, such as color descriptor, shape descriptors and texture descriptors. The used and improved features should be efficient and relevant to describe heterogeneous images. A better images description allows to obtain a satisfactory images classification.

Since the Nineties, Support vector machines (SVMs) did not cease arousing the interest of several researcher communities of various expertise fields.

Such as (Schokopf et al., 1999) which was applied SVMs to insulated handwritten figures recognition, and (Osuna et al., 1997) which was applied SVMs to face recognition. In the majority of cases, SVM performance exceeds those of already established traditional models.

So, for classification, SVMs is used in our retrieval system. SVMs originally formulated for two-class classification problems, have been successfully applied to diverse pattern recognition problems and have become in a very short period of time the standard state of-the-art tool. The SVMs, based on the Structural Risk Minimization (SRM), are primarily devised in order to minimize the upper bound of the expected error by optimizing the trade-off between the empirical risk and the model complexity (Burges, 1998). To achieve this, they construct an optimal hyperplane to separate binary class data so that the margin is maximal.

To evaluate this image retrieval system, we compare it with an other Content-based image retrieval (CBIR) system: the QUadtree-based Index for image retrieval and Pattern search (QUIP-tree).

QUIP-tree indexing structure permits to store the visual characteristics of the various areas in the image. Database images are first of all compared globally with the query image. Then, if its global similarity with the query image is lower than a certain similarity threshold, the under-areas of homologous images are compared, so on until reaching the bottom level (Genevire et al., 2004) (Kachouri et al., 2007).

The paper is organized as follows: Section II describes the CBIR system Structure, and the SVM approach. Section III deals with the different features used in our system, and details the basic improvements done. Experimental results, with a brief description of the QUIP-tree technique are presented in section IV. Finally we conclude in section V.

# 2 CBIR SYSTEM

In this section, we first review the CBIR theory and describe its system Structure. Then we briefly outline the SVM classifier, and QUIP-tree technique.

## 2.1 Content-based Image Retrieval

CBIR is today ubiquitous in computer vision. Similarity queries on feature vectors have been widely used to perform content-based retrieval of images. In fact nowadays, CBIR systems allow image access according to their visual characteristics such as color, texture, shape, etc,..., by means of similarity measures. The smaller the similarity distance is, the closer the two images are.

The typical CBIR system architecture, is composed essentially of two stages. The first one is Off Line, where is carried out the feature extraction of each database image, and the storage of each feature in an index database. The second one is On Line, where is carried out the recognition (classification) by computing similarity measures between the query image signature and the index in the corresponding image database.

There are several popular CBIR systems such as: IBMs QUERY-BY-IMAGE-CONTENT (QBIC) which allows to index images using divers features. Visual SEEK (Smith and Chang, 1996) developed by Smith and Chang in the university of COLUMBIA. Surfimage developed in 1995 by INRIA, which is more sophisticated than the other commercial systems. In this paper, we propose a new CBIR system destined for heterogenous image database.

## 2.2 Support Vector Machines

SVM is a supervised classification method. The supervised classification, supposes that there is already an image classification. So it uses necessarily training methods which from images already classified, allow classifying new images. For image indexing systems, supervised classification allows to build a model which will classify as well as possible new images, from a classified image database.

First, in the Off Line stage: we use a training image database, which is represented by visual descriptors. With the labeled training database images, SVM learns a boundary (i.e., hyper plane) separating the relevant images from the irrelevant images with maximum margin. The images on a side of boundary are considered as relevance, and on the other side are looked as irrelevance.

Second, in the On Line stage: using the built model (boundary computed in the first stage), SVM allows to classify an evaluation image database, which must be also represented by visual descriptors.

SVM have recently attracted a lot of researchers from the machine learning and pattern classification community for its fascinating properties such as high generalization performance and globally optimal solution (Burges, 1998). In SVM, original input space is mapped into a higher dimensional feature space in which an optimal separating hyper-plane is constructed on the basis of SRM to maximize the margin between two classes, i.e., generalization ability.

#### 2.2.1 The Separable Case

Given a set of labeled images $(x_1,y_1)$, ..., $(x_n,y_n)$, $x_i$ is the feature representation of one image, $y_i \in \{-1,+1\}$ is the class label ($-1$ denotes negative and +1 denotes positive).
The goal is to find a boundary such as all the elements, with the same annotation, are on the same side. So we must find a vector $w$ and a real $b$ such as:

$$y_i(w.x_i+b) > 0, \forall i \in [1,n] \quad (1)$$

we can take so, a decision function:

$$f(x) = sign(w.x+b) \quad (2)$$

This decision function is invariant by scale change, so we choose to find the boundary which verify $wx+b=\pm 1$ for nearest elements to margin, what amounts minimizing $\|w\|^2$ such as:

$$y_i(w.x_i+b) \geq 1, \forall i \in [1,n] \quad (3)$$

Using the Lagrangian, the problem amounts maximizing $W$ on $\alpha$, and the decision function is written as follows:

$$f(x) = sign(\sum_{i=1}^{n} y_i \alpha_i x.x_i + b) \quad (4)$$

We note that if we omit the sign operator in the decision function, we obtain a belonging measurement to the required category.

#### 2.2.2 The Non Separable Case

The above algorithm for separable data, when applied to non-separable data, will find no feasible solution. So a flexible margin may be introduced, by accepting bad classification for certain elements. This amounts to raising each $\alpha_i$ by a constant C.
Moreover, linear separation is not adapted to all problems, and it is often preferable to introduce a kernel $k(x,x')$ which replaces the scalar product $x.x'$.
The classification function can be written as:

$$f(x) = sign(\sum_{i} \alpha_i y_i . k(x_i,x) + b) \quad (5)$$

#### 2.2.3 Choice of Kernel

The first kernel investigated for the pattern recognition problem were the following:

$$k(x,y) = (x.y+c)^d \quad \text{Polynomial} \quad (6)$$

$$k(x,y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}} \quad \text{Gaussian} \quad (7)$$

$$k(x,y) = tanh(x.y+\theta) \quad \text{Sigmoidal} \quad (8)$$

The most commonly used kernel is the gaussian one. Since it allows to exploit the distance $d$ placed into exponential:

$$k(x,y) = e^{-\frac{d(x-y)^2}{2\sigma^2}}$$

# 3 USED AND IMPROVED FEATURES

Feature (content) extraction is the basis of CBIR. Recent CBIR systems retrieve images based on visual properties.

As we use an heterogeneous image database, images are various categories, and we can find a big difference between their visual properties. So a unique feature or a unique feature kind, cannot be relevant to describe the whole image database. Then in this paper we are interested by divers visual feature extraction such as color, shape, texture.

## 3.1 Color Features

Color is one of the most important image indexing features employed in CBIR because it has been shown to be effective in both the academic and commercial arenas. Some of the popular methods to characterize color information in images are Color average and color histograms.

#### 3.1.1 Color Average

The color average of an image is defined by $\bar{x}$, as follows:

$$\bar{x} = (\bar{R}_{(avg)}, \bar{G}_{(avg)}, \bar{B}_{(avg)})^t \quad (9)$$

where: $\bar{Color}_{(avg)} = \frac{1}{N}\sum_{p=1}^{N} Color(p)$. N is the total number of pixels in the image.

#### 3.1.2 Color Histograms

Color Histograms are useful because they are relatively insensitive to position and orientation changes. So, despite they are so simple, they are the most commonly used color feature representation. We extract this feature just by computing the occurrence of each gray levels for R, G, and B color planes of the image.

## 3.2 Shape Features

Shape is a very important descriptor in image database. Generally, shape descriptor indicate the general aspect of an object, which is its contour.

#### 3.2.1 Invariant Moments

Invariant moments are important shape descriptors in computer vision. They are obtained from quotients and powers of moments. One moment is a sum on all image pixels weighted by polynomials related to the pixel positions.

In 1962, HU derived seven bi-dimensional invariant moments (Hu, 1962).

This moments are invariant to scale, rotation and translation.

#### 3.2.2 Sobel Filter

Sobel filter is used for contour detection. So, it is supposed that the image areas are homogeneous and that the contour can be detected on the basis of gray levels discontinuity.
First, we apply Sobel masks to obtain the directional gradients according to x and y:

$$G_x(i,j) = h_x(i,j) \otimes I(i,j), G_y(i,j) = h_y(i,j) \otimes I(i,j) \quad (10)$$

Where $I(i,j)$ is the image gray level information and $h_x(i,j)$, $h_y(i,j)$ are Sobel masks:

$$h_x(i,j) = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}, h_y(i,j) = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}$$

then, gradient norm is computed as follow:

$$G(i,j) = \sqrt{G_x(i,j)^2 + G_y(i,j)^2} \quad (11)$$

### 3.3 Texture Features

Multiresolution approaches to texture analysis have gained wide acceptance over the years as they effectively describe both local and global information (Julesz et al., 1978). For this we use in this paper the *Wavelet texture features.*

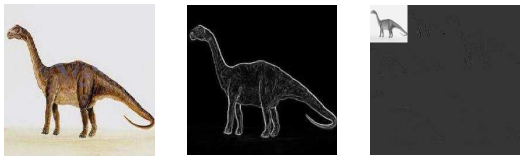
#### 3.3.1 Daubechies Wavelet

Texture features are extracted from Daubechies wavelet coefficients of a two-level decomposition. Daubechies proposed an orthogonal wavelet construction with compact support. Daubechies wavelet has different lengthes called wavelet orders. Daubechies wavelet order, which is always even, is the number of null moments, it is related to the number of oscillations, more there is null moments, more Daubechies wavelet oscillates and so there are more regularities. Indeed, Daubechies wavelet, having M null moments, verify :

$$\Phi(x) = \sqrt{2} \sum_{k=0}^{2M-1} h_{k+1} \Phi(2x-k) \quad (12)$$

$$\Psi(x) = \sqrt{2} \sum_{k=0}^{2M-1} g_{k+1} \Psi(2x-k) \quad (13)$$

with $g_k = (-1)^k . h_{k-1}$, $k = 1,2,...,2M$



a) b) c)

Figure 1: a) Dinosaur, b) Dinosaur gradient norm and c) Dinosaur Daubechies wavelet coefficients of a two-level decomposition.

Wavelet coefficients are $c_{ij}^l(x,y)$, where $l$ is the decomposition level.

Fig. 1 shows Dinosaur image, its gradient norm, and its Daubechies wavelet transformation of a two-level decomposition.

### 3.4 Feature Improvement

To improve the feature size and description, we applied original modifications to some obtained feature coefficients:

#### 3.4.1 Sobel Coefficients

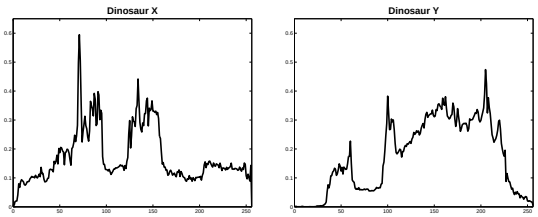
As the coefficient number in the gradient norm is the same as the pixel number in the image, we compute the gradient norm projection according to x and y, in order to reduce this feature size:

$$P_{Xi} = \frac{1}{maxG_{i,j}} \sum_j G(i,j), \text{ and } P_{Yj} = \frac{1}{maxG_{i,j}} \sum_i G(i,j) \quad (14)$$

Despite, this new form is a reduced form of the Sobel feature, it preserves the same properties of the old one.

#### 3.4.2 Moment Coefficients

To obtain more efficient shape description by this feature, we do not use simple moments, which is computed on image pixels, but we compute moments from the gradient norm matrix obtained on sobel feature.




a) b)

Figure 2: New form of Dinosaur Sobel feature: a) The gradient norm projection according to X and b) The gradient norm projection according to Y.

So the particularity of our method, is that it combines Sobel with moments, in a new shape feature description.

#### 3.4.3 Wavelet Coefficients

The lowest frequency coefficients $c^2_{00}(x,y)$ are not inherently useful for texture analysis. Therefore, a direction-independent measure of the high-frequency signal information is obtained by filtering the raw coefficients $c^2_{00}(x,y)$ with the Laplacian.

The texture features are obtained by computing the subband energy of all wavelet coefficients (including the Laplacian filtered $c^2_{00}(x,y)$ coefficients):

$$e^l_{ij} = \frac{1}{MN}\sum_{m=1}^{M}\sum_{n=1}^{N}|c^l_{ij}(m,n)|^2, \qquad (15)$$

where M and N are the dimensions of coefficients $c^l_{ij}(x,y)$. (see Ref. (Serrano et al., 2004) for details).

Table 1: Dinosaur and Rose texture features: subband energy of all Daubechies wavelet coefficients of a two level decomposition.

| Second level decomposition | | | | |
|---|---|---|---|---|
| *Images* | $e^2_{00}$ | $e^2_{01}$ | $e^2_{10}$ | $e^2_{11}$ |
| *Dinosaur* | 226.584 | 11.699 | 8.868 | 6.025 |
| *Rose* | 252.829 | 12.941 | 7.914 | 4.965 |

| First level decomposition | | | |
|---|---|---|---|
| *Images* | $e^1_{01}$ | $e^1_{10}$ | $e^1_{11}$ |
| *Dinosaur* | 5.184 | 3.755 | 2.494 |
| *Rose* | 4.141 | 2.458 | 1.294 |

## 4 EXPERIMENTS

In this section we present, first, a brief description of the QUIP-tree technique, used for comparison purpose. Then we evaluate our proposed system.

### 4.1 Quadtree-based Index for Image Retrieval and Pattern Search

QUIP-tree is an unsupervised classification method. The unsupervised classification, is used when images are not classified. So it is a process by which images are divided into different clusters such as images of the same cluster are as similar as possible and images of different clusters are as dissimilar as possible.

First, in the Off Line stage: we decompose database images into n quadrants, (where n is multiple of four), and we represent them by a visual descriptor by means of quadtree. Then a similarity measure is applied to calculate distance between images. Finally, a clustering of image database is applied.

Second, in the On Line stage: Image query is also decomposed into quadtree structure, after that we compare this query image with image database cluster centers to identify candidate clusters. So query image will be compared, at the end, with only images which belong to candidate clusters to finally find out similar images.

For more details see Ref. (Genevire et al., 2004), (Manouvrier et al., 2005), and (Kachouri et al., 2007).

### 4.2 System Evaluation

For evaluation, we tested our proposed image retrieval system, on an heterogenous image database composed of eight clusters: a collection of 400 images (50 images by cluster). The used heterogeneous database contains images having large difference in colors, shapes, and textures. Some samples are shown in Fig. 3.

To quantitatively evaluate the performances of this system, we have carried out the following tests. Queries representing different clusters are picked from the image database. Then, for each query image, a list of similar images is found in the image database, using SVM classifier.

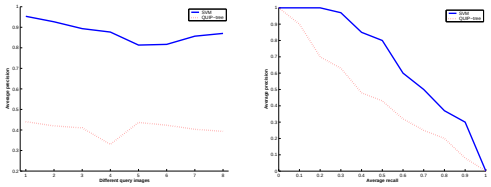
For evaluation purposes, we compare the results of our image retrieval system with other well known classification techniques QUIP-tree (see Fig. 4. $(a)$).

We subsequently computed the retrieval efficiency using the standard retrieval benchmarks: precision and recall (Bimbo, 2001). Let the total number of images retrieved for a query be 50, and let $x1$ be the number of images retrieved that are similar to the query. Let $x2$ be the actual number of images similar to the query in the image database. Evaluation standards recall and precision are defined as follows:

$$precision = \frac{x1}{50}\times 100\%, \text{ and } recall = \frac{x1}{x2}\times 100\% \qquad (16)$$

The criteria of precision and recall are often represented like graphs called precision/recall curves. In these decreasing curves, the precision is represented in terms of recall values. Ideally precision is equal to 1 for all recall values (see Fig. 4. $(b)$).

Figure 3: Samples of the used heterogeneous image database.



a) b)

Figure 4: a) Average precision graph for SVM and QUIP-tree using a combination of color, shape, and texture descriptor and b) Precision/recall curves.

Since QUIP-tree is based on a computation of similarity / dissimilarity, it is efficient, only for small dimensions (only one or two same kind features). So, in (Kachouri et al., 2007), QUIP-tree proved more better than SVMs method in term of recognition rate results according to different image request, because the descriptors used for comparison are simple features (color histogram and color average), which do not permit to build a reliable model of SVMs, and image database used for tests contains synthetic images, where there are only a color variation between the different database images.

But, as soon as dimension is increased, by using more features (in order to improve description), the QUIP-tree retrieval accuracy decreases significantly, from where the favor of SVMs which in such case pass to a larger dimension, using a kernel.

Indeed, by comparing the results of our retrieval system based on SVM classifier with those of QUIP-tree, we find that in all experimental results the SVM retrieval accuracy is better than the QUIP-tree one (as shown in Fig. 4).

Fig. 5 shows the first twelve retrieval results for an example of two query image, using our proposed image retrieval system. The image displayed first is the query and ranking goes from left to right and top to bottom.

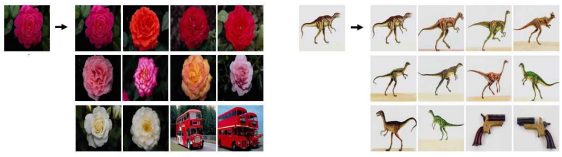


Figure 5: Retrieval results for two query image using our proposed image retrieval system.

## 5 CONCLUSIONS

In this paper, we have presented an heterogeneous image retrieval system based on feature extraction and SVM classifier. To evaluate this system, several kinds of features are used and improved, such as color, shape, and texture features.

The improved features have allowed obtaining a satisfactory image description. The relevance of this description is tested through an SVM classifier. A comparison with QUIP-tree technique is carried out.

As we use a real heterogenous image database, and several kinds of features to indexing images, SVMs prove more better than QUIP-tree method in term of retrieval accuracy and precision/recall curves.

Moreover, in QUIP-tree method, we calculate all distances between each image request and the other database images; whereas, with SVMs, once the model is built, each image request will be just evaluated. So, for consequent database images the SVMs answer is faster than the QUIP-tree one.

The obtained results show that the proposed system provides good accuracy recognition.

## REFERENCES


Bimbo, A. (2001). Visual information retrieval. *Morgan Kaufmann Publishers*.

Burges, C. (1998). A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discovery*, 2(2):121–167.

Genevire, J., Maude, M., Vincent, O., and Marta, R. (2004). Indexation multi-niveau pour la recherche globale et partielle d'images par le contenu. In *BDA*.

Hu, M. (1962). Visual pattern recognition by moment invariants. *IEEE Transactions information Theory*, 8:179–187.

Julesz, B., Gilbert, E., and Victor, J. (1978). Visual discrimination of textures with identical third-order statistics. *Biol. Cybern.*, 31:137–140.

Kachouri, R., Djemal, K., Sellami-Masmoudi, D., Maaref, H., and Derbel, N. (2007). On the heterogeneous image retrieval with quip-tree. In *SSD*.

Manouvrier, M., Rukoz, M., and Jomier, G. (2005). *Spatial Databases : Technologies, Techniques and Trend, Quadtree-Based Image Representation and Retrieval*, chapter 4, pages 81–106. IDEA Group Publishing, Information Science Publishing and IRM Press.

Osuna, E., Freund, R., and Girosi, F. (1997). Training support vector machines: an application to face detection.

Schokopf, B., Burges, C., and Smola, A. (1999). *Introduction to support vector learning*, chapter 1. Advances in Kernel Methods - Support Vector Learning.

Serrano, N., Savakisb, A., and Luoc, J. (2004). Improved scene classification using fficient low-level features and semantic cues. *Pattern Recognition*, 37:1773–1784.

Smith, J. and Chang, S. (1996). Tools and techniques for colour image retrieval. In *IS T/SPIE Proceedings*, volume 2670, pages 426–437, San Jose, CA, USA.

# IDENTIFICATION OF MULTI-DIMENSIONAL SYSTEM BASED ON A NOVEL CRITERION

Yue Zhao, Kueiming Lo
*School of Software, Tsinghua University, Key Lab for ISS, MOE China, Tsinghua University, Beijing 100084, P. R. China*
*yue-zhao07@mails.tsinghua.edu.cn, gluo@tsinghua.edu.cn*

Wook-Hyun Kwon
*School of Electrical Engineering, Seoul National University, Seoul 151-742, Korea*
*whkwon@cisl.snu.ac.kr*

Keywords: Multi-dimensional system, identification criterion, ARMAX model, recursive algorithm.

Abstract: Most system recursive identification algorithms are based on the prediction error (PE) criterion. Such a recursive algorithm only considers the present estimation residual error instead of all estimation residuals. It would result in large estimation error when the signal noise disturbs strongly. In this paper, a new identification criterion is proposed. It considers both the errors between the actual outputs and the estimation result and the difference of each estimation error. Under this criterion, a new recursive algorithm MSDCN (Multi-dimensional System Disturbed by Color Noise) is proposed. For multi-dimensional systems, weighting different values on the estimation errors and the difference of each error, MSDCN could both decrease the estimation errors and got smooth prediction curves. Several simulation examples are given to illustrate the method's anti-disturbance performance.

## 1 INTRODUCTION

There already have many recursive algorithms for modeling systems disturbed by white noises (Andersson and Broman 1998, Ljung 1999, Griffith Jr 1999, Mershed 2000, etc.). The most characteristic feature of disturbance, however, is that its value is not known beforehand. A physical system is affected by many factors such as color noise disturbance and an un-modeled structure. These types of algorithms encounter difficulties when the measurements are disturbed (Kuo 2000, Trump 2001). For multi-discrete systems disturbed by color noise, current identification algorithms cannot give precise estimate results either (Schoukens 1991, Ljung 1985).

Efficiency modeling depends on the choice of identification criterion. Combining frequency-domain method and time-domain method the GPE criterion was proposed for interference systems (Lo and Kwon 2002, 2003). Some extended recursive algorithms (ERA) were put forward based on the GPE criterion. (Lo and Kimura 2003, Lo *et al.* 2006, and Lo and Huang 2006) Usually, a GPE criterion contains a weighting matrix, in which there are many parameters for free choice. When the weighting matrix is the identity, the ERA becomes the recursive least squares algorithm (Lo and Kimura 2003). In addition, the standard Yule-Walker method uses identity matrix as weighting matrix. It may have poor accuracy, and increasing the dimension of the weighting matrix may well degrade the accuracy (Stoica, Friedlander and Soderstrom 1987). Stoica and Jansson (2001) proposed another method which derived the optimal weight in a simple way and guaranteed the optimal weighting matrix to be consistent and non-negative definite, while still the choice of weighing matrix is hard. Furthermore, for all the above raised implementations, a positive definite weighting matrix must be weighted out in order to get a reliable estimate. The optimal weight in general depends on unknown quantities and hence must be itself estimated before its use become possible.

Considering both the prediction errors and the difference of each error, performance criterion in this paper applies different weights. One part is the errors between the actual outputs and the other is the estimation result and the difference of each estimation error. Further, this paper develops a recursive weighting matrix for fast calculation in recursive algorithms. In this matrix, the values of each element depends on the

weights of the performance function. Based on this new performance function, a new recursive algorithm for multi-system identification, MSDCN, is proposed.

MSDCN algorithm is new not only because it is based on a new performance function, but also because it is based on a new concept of estimating the noise part separately. Its two-step estimation feature also make it a novel method. The MSDCN algorithm has good anti-disturbance. For multi-dimensional system, using the novel criterion to do the complex estimation in each dimension of parameters, MSDCN can give more precise results than current algorithms do.

In the second part of this paper, the extended recursive algorithm for multi-dimensional systems is introduced; The third part proposes a novel identification criterion; In the fourth part a recursive algorithm for multi-dimensional system, MSDCN, is proposed; An analysis of the performance of new criterion and new algorithm is given in the fifth section, and then its simulation results are compared with other algorithms.

## 2 EXTENDED RECURSIVE ALGORITHM FOR MULTI-DIMENSIONAL SYSTEM

Consider system:

$$A(q)y(t) = B(q)u(t) + w(t) \tag{1}$$

where, $y(t)$ and $w(t)$ are $p$-dimensional vectors, $u(t)$ is $m$-dimensional vector. $y(t)$, $u(t)$, and $w(t)$ are system output, input, and noise respectively. $A(q)$, $B(q)$ are the backward operator $q^{-1}$ polynomial expression

$$A(q) = I + \sum_{k=1}^{n_a} A_k q^{-k}, \quad B(q) = \sum_{k=1}^{n_b} B_k q^{-k}.$$

Denote:

$$\begin{aligned} \theta &= (A_1, A_2, \cdots, A_{n_a}, B_1, B_2, \cdots, B_{n_b})^T \\ \varphi_t &= (-y^T(t-1), \cdots, -y^T(t-n_a), \\ &\quad u^T(t-1), \cdots, u^T(t-n_b))^T \end{aligned}$$

where $\theta \in R^{(n_a \cdot p + n_b \cdot m) \times p}$ is a matrix formed by the system part parameters. $\varphi_t \in R^{n_a \cdot p + n_b \cdot m}$ is a regression vector. Then (1) is rewritten as

$$y(t) = \theta^T \varphi_t + w(t) \tag{2}$$

Denote $\varepsilon_t$ be the prediction error of system output y(t):

$$\varepsilon_t = y(t) - \theta^T \varphi_t, \tag{3}$$

Then the identification criterion can be expressed as:

$$J(N) = tr([\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_N]^T Q(N) [\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_N]) \tag{4}$$

where $trA$ represents the trace of matrix A. The weighting matrix $Q(N) \in \mathbf{R}^{N \times N}$ is a symmetrical positive definite matrix and expressed as:

$$Q(N) = \begin{pmatrix} Q(N-1) & \alpha(N) \\ \alpha(N)^T & q_N \end{pmatrix}, \quad t = 1, 2, \cdots, N$$

where $\alpha(N) \in \mathbf{R}^{N-1}$. Denote:

$$\Phi(N) = (\varphi_1, \varphi_2, \cdots, \varphi_N)^T \quad Y(N) = (y_1, y_2, \cdots, y_N)^T$$

Then the identification criterion (4) can be expressed as:

$$J(N) = tr([Y(N) - \Phi(N)\theta]^T Q(N) [Y(N) - \Phi(N)\theta] \tag{5}$$

Since

$$\begin{aligned} J(N) &= tr(Y^T(N)Q(N)Y(N)) - tr(\theta^T \Phi^T(N)Q(N)Y(N)) \\ &\quad - tr(Y^T(N)Q(N)\Phi(N)\theta) + tr(\theta^T \Phi^T(N)Q(N)\Phi(N)\theta) \end{aligned}$$

The gradient of J(N) is:

$$\begin{aligned} &\frac{\partial J(N)}{\partial \theta} = \frac{\partial}{\partial \theta}[tr(Y^T(N)Q(N)Y(N)) \\ &- tr(\theta^T \Phi^T(N)Q(N)Y(N)) \\ &- tr(Y^T(N)Q(N)\Phi(N)\theta) \\ &+ tr(\theta^T \Phi^T(N)Q(N)\Phi(N)\theta(N))] \\ &= \frac{\partial}{\partial \theta}[tr(Y^T(N)Q(N)Y(N))] - \\ &\frac{\partial}{\partial \theta}[tr(\theta^T \Phi^T(N)Q(N)Y(N))] \\ &- \frac{\partial}{\partial \theta}[tr(Y^T(N)Q(N)\Phi(N)\theta)] \\ &+ \frac{\partial}{\partial \theta}[tr(\theta^T \Phi^T(N)Q(N)\Phi(N)\theta(N))] \\ &= -(Y^T(N)Q(N)\Phi(N))^T - \Phi^T(N)Q(N)Y(N) + \\ &(\Phi^T(N)Q(N)\Phi(N) + \Phi^T(N)Q(N)^T\Phi(N))\theta(N) \end{aligned}$$

and

$$\frac{\partial J^2(N)}{\partial \theta^2} = \Phi^T(N)Q(N)^T\Phi(N)$$

where $Q(N) = Q^T(N)$ and $\frac{\partial J^2(N)}{\partial \theta^2}$ is positive and definite. Let

$$\frac{\partial J(N)}{\partial \theta} = 0$$

which minimizes $J(N)$ and yields:

$$\theta(N) = [\Phi(N)^T Q(N) \Phi(N)]^{-1} \Phi(N)^T Q(N) Y(N) \tag{6}$$

$t = 1, 2, \cdots$ Then at time t, denote:

$$\begin{aligned} P_t &= \Phi_t^T Q_t \Phi_t \\ a_t &= 1 + \varphi_t^T P_{t-1}^{-1} \Phi_{t-1}^T \alpha_t \\ \sigma_t &= q_t - \alpha_t^T \Phi_{t-1} P_{t-1}^{-1} \Phi_{t-1}^T \alpha_t \\ b_t &= a_t + a_t^{-1} \sigma_t \varphi_t^T P_{t-1}^{-1} \varphi_t \end{aligned}$$

Then we can get:

**Theorem 1**

$$\begin{cases} P_t^{-1} = & P_{t-1}^{-1} \\ & - b_t P_{t-1}^{-1} (\varphi_t \beta_t^T \Phi_{t-1} + \Phi_{t-1}^T \beta_t \varphi_t^T) P_{t-1}^{-1} \\ & + \frac{1}{a_t b_t} P_{t-1}^{-1} (\varphi_t^T P_{t-1}^{-1} \varphi_t \Phi_{t-1}^T \beta_t \beta_t^T \Phi_{t-1} \\ & - \sigma_t \varphi_t \varphi_t^T) P_{t-1}^{-1} \\ \theta_t = & \theta_{t-1} \\ & + \frac{1}{a_t b_t} P_t^{-1} (\beta_t \Phi_t^T \beta_t + \sigma_t \varphi_t)(y_t - \theta_{t-1}^T \varphi_t) \\ & + \frac{1}{a_t b_t} P_{t-1}^{-1} (\beta_t \varphi_t - \varphi_t^T P_{t-1}^{-1} \varphi_{t-1}^T \beta_t) \beta_t^T (Y_{t-1} \\ & - \Phi_{t-1} \theta_{t-1}) \end{cases}$$

As Extended Recursive Algorithm for multi-dimensional system.

## 3 IDENTIFICATION CRITERION

New identification criterion is:

$$J(N) = \lambda \sum_{t=1}^{N} \| \varepsilon_t \|^2 + \mu \sum_{t=1}^{N-1} \| \varepsilon_{t+1} - \varepsilon_t \|^2 \quad (7)$$

where $\varepsilon_t$ represents the estimation errors at time t, in multi-dimensional system, $\varepsilon_t$ is p-dimensional vector, which dimension is the same as the output y(t).

$\lambda$ represents the weight on the estimation errors

$\mu$ represents the weight on the difference between each estimation error.

Transforming equation (7), we can get

$$\begin{aligned} J(N) &= \sum_{t=1}^{N} (\lambda \varepsilon_t^T \varepsilon_t) + \sum_{t=1}^{N-1} \mu (\varepsilon_{t+1} - \varepsilon_t)^T (\varepsilon_{t+1} - \varepsilon_t)) \\ &= \sum_{t=1}^{N} tr(\lambda \varepsilon_t \varepsilon_t^T) + \sum_{t=1}^{N-1} \mu tr(\varepsilon_{t+1} - \varepsilon_t)(\varepsilon_{t+1} - \varepsilon_t)^T \\ &= tr \left( (\lambda + 2\mu) \sum_{t=1}^{N} \varepsilon_t \varepsilon_t^T - \mu \sum_{t=1}^{N-1} \varepsilon_{t+1} \varepsilon_t^T - \mu \sum_{t=1}^{N-1} \varepsilon_t \varepsilon_{t+1}^T \right) \end{aligned}$$

Assume $E(N) = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_N)^T \in N \times p$ and

$$Q(N) = \begin{pmatrix} q_{11} & q_{12} & \cdots & q_{1N} \\ q_{21} & q_{22} & \cdots & q_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ q_{t1} & q_{t2} & \cdots & q_{NN} \end{pmatrix} \in N \times N,$$

in which,

$$\begin{cases} q_{i,i} = \lambda + 2\mu \\ q_{i+1,i} = -\mu \\ q_{i,i+1} = -\mu \end{cases} \qquad i = 1, 2, \cdots, N.$$

Then equivalent with equation (3), we can get

$$J(N) = trE(N)^T Q(N) E(N) \quad (8)$$

as the new expression for the performance criterion.

Thus at time t, the weighting matrix $Q_t$ can be expressed in this recursive form:

$$Q_t = \begin{pmatrix} Q_{t-1} & \beta_t \\ \beta_t^T & p_t \end{pmatrix}, \qquad t = 1, 2, \cdots, N$$

in which, $\beta_t = [0, 0, \cdots, -\mu] \in \mathbf{R}^{t-1}$ and $p_t = \lambda + 2\mu$.

## 4 RECURSIVE ALGORITHM FOR MULTI-DIMENSIONAL SYSTEM

Consider the ARMAX model

$$A(q)y(t) = B(q)u(t) + C(q)w(t) \quad (9)$$

where, $y(t)$ and $w(t)$ are p-dimensional vectors, $u(t)$ is m-dimensional vector. $y(t)$, $u(t)$, and $w(t)$ are system output, input, and color noise respectively. $A(q)$, $B(q)$, $C(q)$is the backward operator $q^{-1}$ polynomial expression

$$A(q) = I + \sum_{k=1}^{n_a} A_k q^{-k}, \quad B(q) = \sum_{k=1}^{n_b} B_k q^{-k}, \quad C(q) = I + \sum_{k=1}^{n_c} C_k q^{-k}.$$

where $A_k, C_k \in p \times p$, and $B_k \in p \times m$. Denote:

$$\theta = (A_1, A_2, \cdots, A_{n_a}, B_1, B_2, \cdots, B_{n_b})^T,$$

$$\varphi_t = (-y^T(t-1), \cdots, -y^T(t-n_a), u^T(t-1), \cdots, u^T(t-n_b))^T,$$

$$\rho = (C_1, C_2, \cdots, C_{n_c})^T,$$

$$\phi_t = (w(t-1), w(t-2), \cdots, w(t-n_c))^T.$$

in which $\theta \in R^{[n_a \cdot p + n_b \cdot m] \times p}$ is the matrix formed by the system part parameters; and $\rho \in R^{[n_c \cdot p] \times p}$ is the matrix formed by the system's noise part parameters.

Here we introduce a two-step algorithm to do the estimation for system (9).

Step 1: Transform (9) into:

$$y(t) - \rho_{t-1}^T \phi_t = \theta_t^T \varphi_t + w(t) \quad (10)$$

Estimate the system parameters θfor system (10). For the identification of the parameters of system θ, use the following measurements:

$$J(t) = tr(\lambda \sum_{t=1}^{N} \| \varepsilon_t \|^2 + \mu \sum_{t=1}^{N-1} \| \varepsilon_{t+1} - \varepsilon_t \|^2). \quad (11)$$

Step 2: Transform (10) into:

$$y(t) - \theta_t^T \varphi_t = \rho_t^T \phi_t + w(t). \quad (12)$$

Estimate the parameter ρ of filter $C(q)$ of system (9).

Denote

$$\Phi_t = (\varphi_1, \varphi_2, \cdots, \varphi_t)^T, Y_t = (y_1, y_2, \cdots, y_t)^T,$$
$$g_t = y_t - \rho_{t-1}^T \phi_t, G_t = (g_1, g_2, \cdots, g_t)^T$$

And according to equation (8), the novel criterion $J(t)$ can be expressed as:

$$\begin{aligned} J(t) &= trE(t)^T Q_t E(t) \\ &= tr[\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_t]^T Q_t [\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_t] \\ &= tr[G_t - \Phi_t \theta]^T Q_t [G_t - \Phi_t \theta] \end{aligned}$$

If $\Phi_t^T Q_t \Phi_t$ is not singularthen minimize $J_t$ to get the system's parameter vector $\theta_t$'s optimal solution:

$$\theta_t = [\Phi_t^T Q_t \Phi_t]^{-1} \Phi_t^T Q_t G_t, \quad t = 1, 2, \cdots, N. \quad (13)$$

Similar with Extended Recursive Algorithm, we can get MSDCN algorithm for ARMAX model (9):

$$\left\{ \begin{array}{rl} \theta_M(t) = & \theta_M(t-1) \\ & + \frac{1}{a_t b_t} P_t^{-1} (\beta_t \Phi_t^T \beta_t + \sigma_t \varphi_t)(y_t \\ & - \rho_M(t-1)^T \phi_t - \theta_M(t-1)^T \varphi_t) \\ & + \frac{1}{a_t b_t} P_{t-1}^{-1} (\beta_t \varphi_t - \varphi_t^T P_{t-1}^{-1} \varphi_{t-1}^T \beta_t) \beta_t^T (G_{t-1} \\ & - \Phi_{t-1} \theta_M(t-1)) \\ P_t^{-1} = & P_{t-1}^{-1} \\ & - b_t P_{t-1}^{-1} (\varphi_t \beta_t^T \Phi_{t-1} + \Phi_{t-1}^T \beta_t \varphi_t^T) P_{t-1}^{-1} \\ & + \frac{1}{a_t b_t} P_{t-1}^{-1} (\varphi_t^T P_{t-1}^{-1} \varphi_t \Phi_{t-1}^T \beta_t \beta_t^T \Phi_{t-1} \\ & - \sigma_t \varphi_t \varphi_t^T) P_{t-1}^{-1} \\ \rho_M(t) = & \rho_M(t-1) + \frac{R_{t-1}^{-1} \phi_t}{1 + \phi_t^T R_{t-1} \phi_t} (y_t \\ & - \theta_M(t)^T \varphi_t - \rho_M(t-1)^T \phi_t) \\ R_t = & R_{t-1} + \phi_t \phi_t^T \\ \widehat{w}(t) = & y(t) - \rho_M(t)^T \phi_t - \theta_M(t)^T \varphi_t \end{array} \right. \quad (14)$$

in which,

$$\left\{ \begin{array}{l} p_t = \lambda + 2\mu \\ \beta_t = [0, 0, \cdots, -\mu] \in \mathbf{R}^{t-1} \end{array} \right.$$

Initially, θ and ρ can be zero matrix; $R_0$ , $P_0$ together constitute the identity matrix.

# 5 SIMULATIONS

All the simulations were conducted in the same computation environment. The main criterions of the computer were: CPU 1.66GHz, RAM 1G bytes and with Windows XP OS.

**Experiment.** A Black-box model is as follows:

$$y(t) = \frac{B_1 q^{-1}}{I + A_1 q^{-1}} u(t) + \frac{I + C_1 q^{-1}}{I + A_1 q^{-1}} w(t)$$

The real parameters of the system were:

$$A_1 = \begin{pmatrix} 0.325 & -1 \\ 0.5 & -1.1 \end{pmatrix},$$

$$B_1 = \begin{pmatrix} 1.9 & 3.7 \\ -0.4 & -0.8 \end{pmatrix},$$

$$C_1 = \begin{pmatrix} -0.7 & -0.1 \\ 0.8 & 0.9 \end{pmatrix}.$$

The color noise sequence $\{w(t)\}_1^N$ was a two dimensional vector, of which each dimension was composed of different sawtooth wave and random number generator with variance 2 and the input signal $u(t) = [1\ 2]$, each dimension of which was generated by a square generator. The experiment was conducted using the LS method and the algorithm MSDCN method with the sample number $N = 1000$. The results are shown in Figures 1,2,3,4,5,6 and the statistics are as follows:

(1)Least-Squares method

The system statistics results are:

$$\theta = \begin{pmatrix} 0.2889 & 0.0361 \\ -0.9823 & -0.0177 \\ 0.3855 & 0.1145 \\ -1.2366 & 0.1366 \\ 1.7990 & 0.1010 \\ 3.5981 & 0.1019 \\ -0.1692 & -0.2308 \\ -0.3384 & -0.4616 \\ -0.7499 & 0.0499 \\ 0.0907 & -0.1907 \\ 0.9270 & -0.1270 \\ 0.8506 & 0.0494 \end{pmatrix}.$$

The simulation results are shown in Figure 1, 2 and 3.

(2)MSDCN method

if we choose $\lambda = 0.8 and \mu = 0.2$,

then according to (7), $\beta_t = [0, \cdots, -0.2, -0.2]$ and $p_t = 1.2$.

Then the system statistics results are:

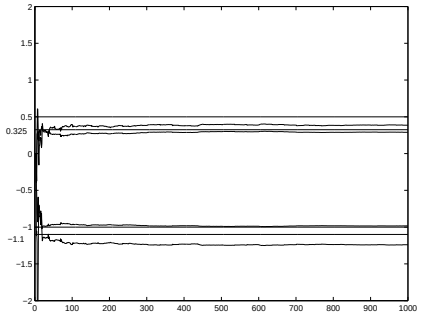


Figure 1: Estimation results of parameter A under ELS.

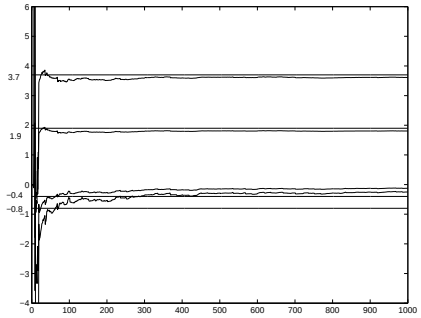


Figure 2: Estimation results of parameter B under ELS.

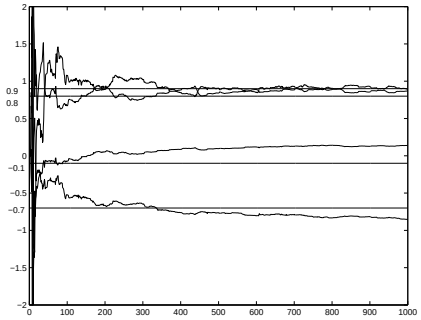


Figure 3: Estimation results of parameter C under ELS.

$$\theta = \begin{pmatrix} 0.2620 & 0.0630 \\ -0.9602 & -0.0398 \\ 0.4448 & 0.0552 \\ -1.2425 & 0.1425 \\ 1.7595 & 0.1405 \\ 3.5189 & 0.1811 \\ -0.4267 & 0.0267 \\ -0.8534 & 0.0534 \\ -0.6499 & -0.0501 \\ 0.0409 & -0.1409 \\ 0.8232 & -0.0232 \\ 0.9362 & -0.0362 \end{pmatrix}$$

The simulation results are shown in Figure 4, 5 and 6.

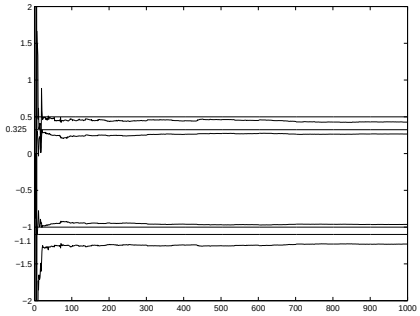


Figure 4: Estimation results of parameter A under MSDCN.

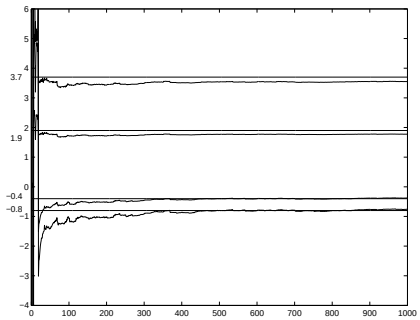


Figure 5: Estimation results of parameter B under MSDCN.

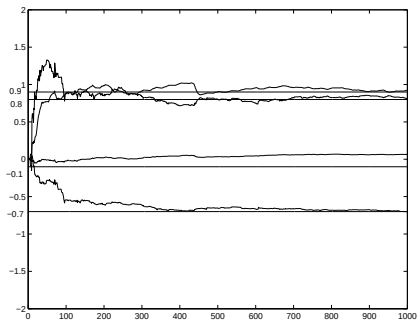


Figure 6: Estimation results of parameter C under MSDCN.

Through a comparison of the estimation results of ELS and MSDCN, we can see that the new performance function has proved to be efficient for the multi-dimensional systems. For the example given above, we weight 0.8 on the estimation errors and 0.2 on the difference between each estimation error, which means we want the estimation curves close to the actual values more than make the curves smooth. The weighting matrix is definite after $\beta_t$ and $p_t$ are fixed by $\lambda$ and $\mu$. Weighting different on $\lambda$ and $\mu$ will result in different estimation results. Those figures show that MSDCN has both decreased the estimation errors and got smooth prediction curves.

## 6 CONCLUSIONS

This paper proposes a new identification criterion for multi-dimensional system disturbed by color noise, and further develops a recursive algorithm, MSDCN, based on it. Weighting both on the prediction errors and the difference of each prediction error, the identification criterion makes the weighting matrix definite in calculation. Based on this performance criterion, MSDCN is developed using a two-step method to estimate both the system parameters and the noise part. The MSDCN algorithm has high anti-disturbance performance in the prediction of multi-dimensional systems disturbed by color noise. It both decreased the estimation errors and got smooth prediction curves. The performance of the MSDCN algorithm was demonstrated by simulations.

## ACKNOWLEDGEMENTS

This work is supported by the Funds NSFC60672110, NSFC60474026,and the JSPS Foundation.

## REFERENCES

Andersson, A. and Broman, H. (1998), A second-order recursive algorithm with applications to adaptive filtering and subspace tracking. *IEEE Trans. Signal Processing*, vol. 46, pp. 1720C1725, June.

Griffith Jr, D. W. and Arce, G. R. (1999), A partially decoupled RLS algorithm for volterra filters. *IEEE Trans. Signal Processing*, vol. 47, pp. 579C582, Feb.

Kuo, C. J. , Jr, R. D. , Lin, C. Y. and Tsai, Y. C. (2000), Set-theoretic estimation based on a priori knowledge of the noise distribution. *IEEE Trans. Signal Processing*, vol. 48, pp. 2150C2156, July.

Lo, K. M. , *et al*. (2006), Empirical frequency-domain optimal parameter estimate for Black-box processes. *IEEE Transactions on Circuits and Systems-I: Regular Papers*, vol.53, No.2, 419-430.

Lo, K. M. , Kwon, W. H. (2003), New identification approaches for disturbed models. *Automatica*.vol.39, No.9, 1627-1634.

Lo, K. M. , Kimura, H. (2003), Recursive Estimation Methods for Discrete Systems. *IEEE Trans. On Circruits and Systems*.vol.49, NO.6, No.6, 439-446.

Ljung, L. (1985), On the estimation of transfer function. *Automatica*vol. 21, 677-696.

Ljung, L. (1999), System Identification: Theory for the User. *Upper Saddle River,NJ: Prentice-Hall*, 1999.

Mershed, R. and Sayed, A. J. (2000), Order-recursive RLS Laguerre adaptive filtering. *IEEE Trans. Signal Processing*, vol. 48, pp. 3000C3010, July.

Stoica, P., Friedlander, B.and Soderstrom, T.(1987) Optimal Intrumental Variable Multistep Algorithms for Estimation of the AR Parameters of an ARMA Process, *Int. J. Control* 45(1987), 2083-2107.

Stoica, P. and Jansson, M.(2001) Estimating Optimal Weights for Instrumental Variable Methods. *Digital Signal Processing - A Review Journal* vol. 11, no. 3, pp. 252-268, Jul. 2001.

Trump, T. (2001), Maximum likelihood trend estimation in exponential noise. *IEEE Trans. Signal Processing*, vol. 49, pp. 2087C2095, Sept.

# SYNTHESIS METHOD OF A PN CONTROLLER USING FORBIDDEN TRANSITIONS SEQUENCES

R. Bekrar, N. Messai, N. Essounbouli, A. Hamzaoui and B. Riera
*CReSTIC, IUT de Troyes, 9 rue de Qubec BP 396, 10026 Troyes Cedex, France*
{*rebiha.bekrar, nadhir.messai, najib.essounbouli, abdelaziz.hamzaoui, bernard.riera*}*@univ-reims.fr*

Keywords: Forbidden state problem, Forbidden transitions sequences, Supervisory control, Discrete event systems, Petri Nets.

Abstract: In this paper, we propose a control synthesis method for Discrete Event Systems (DES) modelled by a bounded ordinary Petri Nets (PN) to treat a forbidden state problem. The considered PN is transitions controllable and contains measurable and non measurable places. The PN controller is synthesised using the forbidden transitions sequences. The latter, are deduced from the PN reachability graph and considered as a forbidden language generated by the PN model. To show the efficiency of the proposed method an illustrative example is presented.

## 1 INTRODUCTION

In Discrete Event Systems (DES), it is important to prevent the system to reach undesirable states. Hence, given a DES model and a specification of the desired behaviours, one must to synthesise an efficient controller in order to achieve this goal. Among tools used to synthesis a DES controller, Petri Nets (PN) have been successfully considered as an efficient formalism for DES control.

This paper is considered as the second phase of our work (Bekrar et al., 2006a; Bekrar et al., 2006b) on the identification of a DES. It allows to complete the identified model by adding control places. Indeed, the identified PN model can generate all the DES states. However, it can generates, sometimes, some states which are not observed during the system functioning. Thus, our objective is to synthesise a PN controller to prevent the reachability of the forbidden states and guarantees the desired ones.

The synthesis problem of PN controllers has been widely treated in the literature. It was introduced either as a forbidden states problem (FSP) or a forbidden state transitions problem (FSTP). The synthesis methods of a PN controller allow to add a set of control places to the initial model in order to generate the desired behaviours. Several approaches are suggested in order to solve this problem. Among them we can find, the logical predicate based solutions (Krogh and Holloway, 1991; Boel et al., 1995; Holloway et al., 1996). The key idea is to perform an off-line structural analysis for determining algebraic expressions. The latter will be evaluated on-line for the current marking in order to decide the firing of the controllable transitions. These approaches are very effective for the on-line PN control. Nevertheless, they are valid only with a specific classes of PN. Other approaches, which give generally nonmaximally permissive solution, have been proposed in the literature (Giua et al., 1992; Moody and Antsaklis, 2000; Basile et al., 2006). They take the form of PN which simplifies the analysis and the implementation of the controller. Based on the theory of regions many works have been developed (Ghaffari et al., 2002; Ghaffari et al., 2003; Achour and Rezg, 2006; Lee et al., 2006). The objective is to solve optimally the supervisory control problem using formal and algebraic characterisations. Finally, the PN controller synthesis problem has been treated as integer linear programming problem (Giua and Xie, 2004; Basile et al., 2007a; Basile et al., 2007b). Nevertheless, these approaches require a very high time computing. Although the synthesis problem of PN controller has been widely treated in the literature, the existing approaches cannot be exploited directly in our case because, it is impossible to characterise the forbidden states by constraints. Indeed, the control specification is mainly important to synthesise a controller. However, we have neither the control specification nor the system description because the identified

PN that we want to control is obtained using only the measurable inputs and outputs system signals. Let's note that, we can determine the forbidden states uniquely by comparing the observed system's states and states reachable by the identified PN model.

This paper deals with the supervisory control problem of DES characterised by forbidden states and modelled by bounded ordinary PN. The PN herein considered is transitions controllable and contains measurable and non measurable places. The latter represent the non measurable outputs system signals. Indeed, the fact that all transitions are controllable then, the PN reachability graph does not contain dangerous markings. In addition, we suppose that we have not the control specification and the initial marking is the unique marked state. Finally, we consider that the PN reachability graph does not contain deadlock states because the input and the output signals observed during the identification phase represent only the normal behaviours of the considered system.
The proposed PN controller is synthesised using the forbidden transitions sequences deduced from the PN reachability graph. The designed PN controller permits to avoid the occurrence of forbidden markings, generated by the non controlled PN model, and guarantees a set of desired behaviours.

In this paper, we start by presenting the considered FSP and defining the control specifications. Then, we present a procedure to calculate the forbidden transitions sequences from the PN reachability graph and introduce a new algorithm to design the PN controller using these sequences. At last, we illustrate the proposed algorithm by an explicative example.

# 2 CONTROL SPECIFICATIONS

We consider the basic supervisory control problem to synthesise a PN controller that avoids the occurrence of forbidden states. In this paper, we assume that the PN that we want to control is established using the I/O sequences describing all the system normal behaviours according to the algorithm presented in (Bekrar et al., 2006b). Hence, our objective is to add some control places in order to guarantee that all the behaviours generated by the identified PN model correspond to those generated by the real system.
Note also that, each PN reachable marking $M_i^{'}$ is composed of two parts: $M_i^{'} = \begin{bmatrix} M_{im}^{'} \\ M_{i\bar{m}}^{'} \end{bmatrix}$ where, the first one represents the marking of the measurable places and the second part represents the estimated marking of the non measurable places.

The control specifications to be addressed in this paper are the forbidden states type and the problem herein considered becomes a FSP. We will treat it as a FSTP as proved in (Ghaffari et al., 2002). To solve this problem, we propose an algorithm that consists in: (1) identifying the set of the forbidden markings, (2) determining the set of the equivalent forbidden state transitions, (3) developing a PN controller.

## 2.1 Identification of Forbidden Markings

Contrary to works proposed to treat the FSP where the forbidden states are defined explicitly by constraints, in our case we must calculate them using the behaviours of the real system and those of the PN model. The behaviours of the considered system are represented by the set of its reachable states called $\mathcal{E}$. Those generated by the PN modelling the system are represented by its reachability graph called $R_G^{'}$. Note that, a marking $M_i^{'} \in R_G^{'}$ reachable by the PN that we want to control is said forbidden marking if its measurable part is not equivalent to any state $E_i$ in $\mathcal{E}$ otherwise, it is a legal marking. Hence, the set of the forbidden markings can be obtained using the following procedure:

Input: $\mathcal{E}$ and $R_G^{'}$.
Output: $M_{fr}$ the set of forbidden markings.
Begin

1. Initialise the set of forbidden markings: $M_{fr} := \emptyset$.
2. For each marking $M_i^{'} \in R_G^{'}$ do:
   (a) If there exists a state $E_i \in \mathcal{E}$ such that $M_{im}^{'} = E_i$ then:
   - $M_i^{'}$ is legal marking.

   (b) Else, $M_i^{'}$ is a forbidden marking.
   - Update the set of forbidden markings: $M_{fr} := M_{fr} \cup M_i^{'}$.

End If.
End For.
End.

Once the set of the forbidden markings is obtained, we should determine the set of the transitions leading to or firing from forbidden markings in order to prevent the firing of these transitions.

## 2.2 Forbidden State Transitions

A state transition in the reachability graph of the PN that we want to control, which fires from a marking $M_i^{'} \in R_G^{'}$ and leads to a marking $M_j^{'} \in R_G^{'}$ ($M_i^{'}[t_i > M_j^{'}$), is said forbidden state transition if and only if one of the following conditions is verified:

- (a) $M_i^{'}$ is forbidden marking or,
- (b) $M_j^{'}$ is forbidden marking.

These can be reformulated, using $\mathcal{E}$ and $R_G^{'}$, as follows: a state transition ($M_i^{'} \xrightarrow{t_i} M_j^{'}$) in $R_G^{'}$ is said forbidden iff:

- $\exists E_i \in \mathcal{E}$: $M_{im}^{'} = E_i$ and $\nexists E_j \in \mathcal{E} : M_{jm}^{'} = E_j$ or,
- $\nexists E_i \in \mathcal{E}$: $M_{im}^{'} = E_i$ and $\exists E_j \in \mathcal{E} : M_{jm}^{'} = E_j$ or,
- $\nexists E_i, E_j \in \mathcal{E} : M_{im}^{'} = E_i$ and $M_{jm}^{'} = E_j$.

*Proof*: the occurrence of forbidden marking can be prevented by avoiding the firing of the transition leading to this marking. Moreover, it is clear that a transition firing from a forbidden marking is a forbidden transition. So, each transition leading to or firing from a forbidden marking is considered as a forbidden transition.
Therefore, the forbidden state transitions set can be obtained using the following procedure:

Input: $\mathcal{E}$, $R_G^{'}$ and $M_{fr}$.
Output: $\Psi$ the set of forbidden state transitions.
Begin

1. Initialise the set of forbidden state transitions: $\Psi = \emptyset$.
2. For each state transition ($M_i^{'} \xrightarrow{t_i} M_j^{'}$) $\in R_G^{'}$ such that $M_i^{'}, M_j^{'} \in R_G^{'}$ do:
   - If (a) or (b) is verified then: ($M_i^{'} \xrightarrow{t_i} M_j^{'}$) is a forbidden state transition.
     - Update $\Psi$:($\Psi$:= $\Psi \cup \{(M_i^{'} \xrightarrow{t_i} M_j^{'})\}$).
   - Else, ($M_i^{'} \xrightarrow{t_i} M_j^{'}$) is a legal state transition.

End If.
End For
End.

Once the forbidden state transitions are determined, the forbidden transitions sequences will be calculated and used for designing a PN controller. More details concerning these steps will be presented in the next section.

# 3 THE PROPOSED PN CONTROLLER

We can solve the forbidden state problem considered in this paper by adding control places $\{p_{c1}, ..., p_{ck}\}$ to the initial PN model $(N, M_0^{'})$. These places are defined as follows:
*Definition 1*: A control place $p_{ci}$ of a PN model $(N, M_0^{'})$ is defined by: (i) $M_0^{'}(p_{ci})$: its initial marking, (ii) $Post(p_{ci}, .)$ and $Pre(p_{ci}, .)$: the weighting vectors of the arcs connecting the transitions of $(N, M_0^{'})$ to $p_{ci}$ and connecting $p_{ci}$ to the transitions of $(N, M_0^{'})$ respectively.

*Remark 1*: Since the considered PN is ordinary then, the arcs weighting values are equal to 1.
In order to solve this problem, we propose to use the forbidden transitions sequences, calculated from the reachability graph of the PN that we want to control, to determine the control places to be added. Note that, these sequences have been used by Lee et al., (2006) but with the constraint asynchronous reachability graph.

## 3.1 The Forbidden Transitions Sequences

Before introducing the computing procedure of the forbidden transitions sequences, let us present the following definitions that will be used in the PN controller design algorithm.
*Definition 2*: A transitions sequence $\sigma_i$ is said forbidden if and only if it allows the firing of at least one forbidden transition. Thus, $\sigma_i = t_1, t_2, ..., t_k, t_l, ..., t_f$ is said forbidden iff: $\exists t_i \in \sigma_i :^* M(t_i) \in M_{fr}$ or $M^*(t_i) \in M_{fr}$ where $^*M(t_i)$ and $M^*(t_i)$ represent respectively the input and the output marking of $t_i$ in $R_G^{'}$, otherwise, it is a legal.
The forbidden transitions sequences are calculated using the PN reachability graph as follows:

Input: $R_G^{'}$, $\Psi$.
Output: $S$ the set of forbidden transitions sequences.
Begin

1. Initialise the set of forbidden transitions sequences: $S := \emptyset$.
2. Calculate $S^{'}$ the set of all legal transitions sequences reachable from the initial marking $M_0^{'}$.
3. Calculate the set of the successors transitions of each transitions sequence $\sigma_i$ in $S^{'}$, noted $Suc(\sigma_i)$.
4. For each transition $t_i$ in the set $Suc(\sigma_i)$ do:

(a) If $t_i$ is a forbidden transition then:
- $\sigma_j := \sigma_i t_i$ is a forbidden transitions sequence.
- Else, stop the successors computation procedure of the transitions sequence $\sigma_j$.

(b) Update $S$ ($S := S \cup \{\sigma_j\}$).

End If.
End For

5. Calculate the set of successors transitions of each transitions sequence $\sigma_i$ in $S$.
6. Go to 4.
7. End.

Note that, in our case, each forbidden transitions sequence $\sigma_i$ is composed of two sub-sequences as follows: $\sigma_i = \underbrace{t_1,...,t_k}_{\sigma_i^{''}} | \underbrace{t_l,...,t_f}_{\sigma_i^{in}}$, with $\sigma_i^{''}$ is a legal transitions sub-sequence authorised to be fired from the initial marking $M_0^{'}$ and it can be equal to the empty string $\varepsilon$ (i.e., $\sigma_i^{''} = \varepsilon$). Thus, it means that there exists forbidden states reachable from $M_0^{'}$ after the firing of only one transition. $\sigma_i^{in}$ is a sub-sequence forbidden to firing from $M_k^{'}$ where $M_k^{'}$ is the marking reachable after the firing of $t_k$. This sub-sequence represents the transitions influenced by the firing of the forbidden transition $t_l$. We note by ${}^*\sigma_i^{in}$ and $\sigma_i^{in*}$ the first and the last transition of $\sigma_i^{in}$ respectively. Hence, all the markings reachable from $M_0^{'}$ after the firing of the sub-sequence $\sigma_i^{''}$, transition after transition, are legal markings. However, the markings reachable from $M_k^{'}$ after the firing of the sub-sequence $\sigma_i^{in}$, transition after transition, are forbidden markings as represented in figure 1.

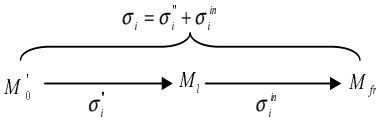



Figure 1: Legal and forbidden markings.

*Definition 3*: Let's $\sigma_i = t_1, t_2, ..., t_k$ transitions sequence. $|\sigma_i|$ designs the length of this sequence, and it represents the number of transitions in this sequence.

*Remark 2*: For each transition $t_i \in \sigma_i$, we note by $|t_i \cap \sigma_i|$ the number of times that the transition $t_i$ appears in $\sigma_i$.

*Definition 4*: Let's $\sigma_i = t_1, t_2, ..., t_k$ a transitions sequence that we consider as a word generated by a PN. The prefix of $\sigma_i$, noted $Pref(\sigma_i)$, is a word $t_1 ... t_j$ with $0 \leq j \leq k$ and $k = |\sigma_i|$. For $q \in \mathbb{N}$: $Pref(\sigma_i)^{(q)}$ represents a prefix of $\sigma_i$ of width equal to $q$.

*Remark 3*: Since PN to be controlled is bounded, then its reachability set is finite. Also, the language generated by this PN is finite prefix-closed language.

## 3.2 Synthesis Algorithm of PN Controller

Based on previous definitions and notations, we propose an algorithm that allows to add control places to the initial PN model for preventing the reachability of forbidden markings. This algorithm contains the following:

1. Construction of the reachability graph of the PN that we want to control.
2. Identification of the forbidden markings and the determination of the forbidden state transitions by browsing the PN reachability graph.
3. Determination of the forbidden transitions sequences set. These latter, we consider them as forbidden words generated by the PN model. Then, we use these sequences to synthesise a PN controller.

Hence, the PN controller will be synthesised using the following algorithm:

Input: $\mathcal{L}$ the system states set, $(N, M_0^{'})$ the PN model.
Output: Controlled PN.
Begin

1. Construct the reachability graph $R_G^{'}$ of the PN model to be controlled $(N, M_0^{'})$.
2. Identify the set of forbidden markings noted $M_{fr}$ as introduced in subsection 2.1.
3. Determine the set of forbidden state transitions $\Psi$ by executing the procedure described in subsection 2.2.
4. Determine the set $S$ of all forbidden transitions sequences firing from the initial marking $M_0^{'}$ by browsing the reachability graph of PN that we want to control as shown in subsection 3.1.
5. Update the set of forbidden transitions sequences $S$ as follows:

(a) For each transitions sequence $\sigma_i \in S$ do:
   i. Calculate the prefix set of this sequence noted $Pref(\sigma_i)$.
   ii. Compare the prefixes of $\sigma_i$ to the prefixes of all the transitions sequences in $S$ as follows:

A. If there exists at least one transitions sequence $\sigma_j$ such that, all the elements of prefixes set of $\sigma_i$ are included in the prefixes set of $\sigma_j$ ($Pref(\sigma_i) \subset Pref()\sigma_j$) then:
   - Eliminate $\sigma_i$from $S$: $S := S\backslash\{\sigma_i\}$

B. Else, $\sigma_i$ still in $S$.

End If
End For.

6. For each forbidden transitions sequence in $S$ do:
   (a) Determine the sub-sequence of legal transitions $\sigma_i^{''}$ and the sub-sequence of forbidden transitions $\sigma_i^{in}$.
   (b) For each sub-sequence of forbidden transitions $\sigma_i^{in}$, we determine ${}^*\sigma_i^{in}$ and $\sigma_i^{in*}$.
   (c) Add a control place $p_{ci}$ as an input of ${}^*\sigma_i^{in}$ and as an output of $\sigma_i^{in*}$.
   (d) Mark $p_{ci}$ with initial marking $M_{c0}(p_{ci}) = |{}^*\sigma_i^{in} \cap \sigma_i^{''}|$. Il there exists several forbidden transitions sub-sequences which have the same ${}^*\sigma_i^{in}$ and $\sigma_i^{in*}$ then, we add one control place $p_{ci}$ and we mark it with initial marking equal to $\max_{i,\sigma_i \in S}(|{}^*\sigma_i^{in} \cap \sigma_i^{''}|)$.

End For.
End.

This algorithm allows to complete the PN model, established using the identification approach proposed previously, by adding control places. The analysis of the algorithm computational complexity shows that it is linear with the number of places, of transitions and, the length of the largest forbidden transitions sequences.

# 4 ILLUSTRATIVE EXAMPLE

To illustrate the proposed algorithm, let us consider the example of a system described by the identified PN model of figure 2:

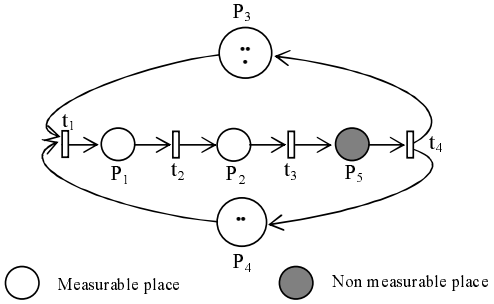



Figure 2: The PN model.

The set of its states is given by: $\mathcal{E} = \left\{ \underbrace{\begin{pmatrix}0\\0\\3\\2\end{pmatrix}}_{E_0}, \underbrace{\begin{pmatrix}1\\0\\2\\1\end{pmatrix}}_{E_1}, \underbrace{\begin{pmatrix}0\\1\\2\\1\end{pmatrix}}_{E_2}, \underbrace{\begin{pmatrix}0\\0\\2\\1\end{pmatrix}}_{E_3}, \underbrace{\begin{pmatrix}2\\0\\1\\0\end{pmatrix}}_{E_4}, \underbrace{\begin{pmatrix}1\\1\\1\\0\end{pmatrix}}_{E_5}, \underbrace{\begin{pmatrix}1\\0\\1\\0\end{pmatrix}}_{E_6} \right\}$.

Firstly, we elaborate the PN reachability graph that is given in figure 3.

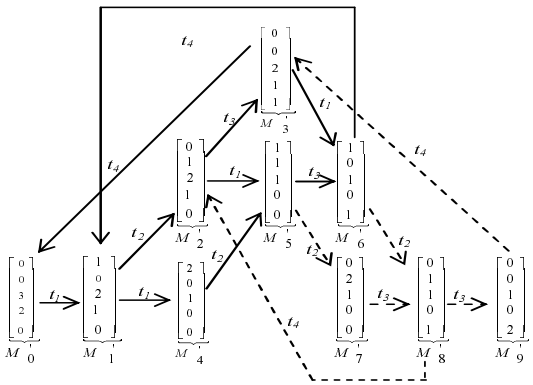



Figure 3: The PN reachability graph.

The set of forbidden markings is: $M_{fr} = \{M_7^{'}, M_8^{'}, M_9^{'}\}$. These markings must be removed from $R_G^{'}$ together with their arcs coming from other markings to these forbidden markings or going from these forbidden markings to other markings in $R_G^{'}$. The set of the equivalent forbidden state transitions are: $\{(M_5^{'} \xrightarrow{t_2} M_7^{'}), (M_7^{'} \xrightarrow{t_3} M_8^{'}), (M_8^{'} \xrightarrow{t_2} M_9^{'}), (M_9^{'} \xrightarrow{t_4} M_3^{'}), (M_6^{'} \xrightarrow{t_2} M_8^{'}), (M_8^{'} \xrightarrow{t_4} M_2^{'})\}$. By applying the procedure described in subsection 4.1, the forbidden transitions sequences firing from the initial marking are: $\sigma_1 = t_1t_2t_3t_1t_2$, $\sigma_2 = t_1t_2t_3t_1t_2t_4$, $\sigma_3 = t_1t_2t_3t_1t_2t_3$, $\sigma_4 = t_1t_2t_3t_1t_2t_3t_4$, $\sigma_5 = t_1t_1t_2t_3t_2$, $\sigma_6 = t_1t_1t_2t_3t_2t_4$, $\sigma_7 = t_1t_1t_2t_3t_2t_3$, $\sigma_8 = t_1t_1t_2t_3t_2t_3t_4$, $\sigma_9 = t_1t_1t_2t_2$, $\sigma_{10} = t_1t_1t_2t_2t_3$, $\sigma_{11} = t_1t_1t_2t_2t_3t_4$, $\sigma_{12} = t_1t_1t_2t_2t_3t_3$, $\sigma_{13} = t_1t_1t_2t_2t_3t_3t_4$. Thus, the set of the forbidden transitions sequences is $S = \{\sigma_{i,i=1,\ldots,13}\}$. To update $S$, we must calculate the prefix of each sequence in $S$. We take as example the sequences $\sigma_1$ and $\sigma_2$: $Pref(\sigma_1) = \{\varepsilon, t_1, t_1t_2, t_1t_2t_3, t_1t_2t_3t_1, t_1t_2t_3t_1t_2\}$, $Pref(\sigma_2) = \{\varepsilon, t_1, t_1t_2, t_1t_2t_3, t_1t_2t_3t_1, t_1t_2t_3t_1t_2, t_1t_2t_3t_1t_2t_4\}$. We note that $Pref(\sigma_1) = Pref(\sigma_2)$ then, we eliminate $\sigma_1$ from $S$: $S := S\backslash\{\sigma_i\}$. By repeating the same process with the remaining sequences, and after updating $S$, we have finally : $S = \{\sigma_2, \sigma_4, \sigma_6, \sigma_8, \sigma_{11}, \sigma_{13}\}$. Then, for each forbidden transitions sequence in $S$ we determine the legal transitions sub-sequence and the forbidden one. We take for example $\sigma_8 = \underbrace{t_1t_1t_2t_3}_{\sigma_8^{''}} | \underbrace{t_2t_3t_4}_{\sigma_8^{in}}$. We remark that $\sigma_8^{''} = t_2$ and

$\sigma_8^{in} = t_4$. Finally, by analysing the remaining forbidden sequences, we see that all the forbidden transitions sub-sequences have the same ${}^*\sigma_i^{in} = t_2$ and $\sigma_i^{in*} = t_4$ for $i \in \{2,4,6,11,13\}$. Therefore, we add one control place $p_{c1}$ with initial marking $M_{oc}(p_{c1}) = \max(|{}^*\sigma_i^{in} \cap \sigma_i^{''}|) = 1$. This place is an input of $t_2$ and an output of $t_4$ as depicted in the figure 4.

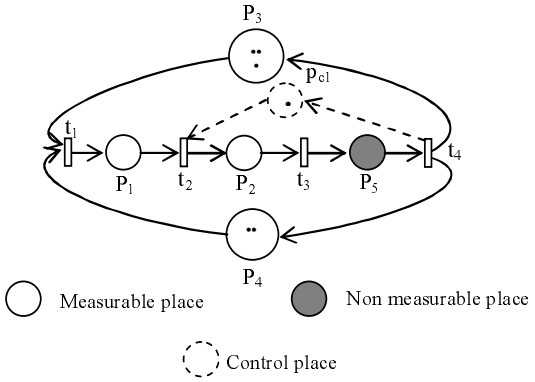



Figure 4: The Controlled PN model.

# 5 CONCLUSIONS

This paper presents a synthesis method of a PN controller to solve a FSP of DES modelled by a bounded ordinary PN. The model to be controlled is transitions controllable. Using the system behaviours and those generated by the considered PN model, forbidden markings are identified and the equivalent forbidden state transitions are determined. Then, the forbidden transitions sequences deduced from the PN reachability graph are used to synthesise a PN controller. The latter is maximally permissive within the specifications that guarantees the desired behaviours.
As Future work, we will generalise this method to PN with uncontrollable transitions.

# REFERENCES

Achour, Z. and Rezg, N. (2006). Time floating general mutual exclusion constraints (tfgmec). In *IMACS, Multiconference on Computational Engineering in Systems Applications*.

Basile, F., Carbone, C., and Chiacchio, P. (2007a). Feedback control logic for backward conflict free choice nets. In *IEEE Transaction on Automation and Control*.

Basile, F., Chiacchio, P., and Giua, A. (2006). Suboptimal supervisory control of Petri nets in presence of uncontrollable transitions via monitor places. In *Automatica*.

Basile, F., P.Chiacchio, and Giua, A. (2007b). An optimisation approach to Petri net monitor design. In *IEEE Transaction on Automation and Control*.

Bekrar, R., Messai, N., Essounbouli, N., Hamzaoui, A., and Riera, B. (2006a). Identification of discrete event systems using ordinary Petri nets. In *IAR-ACD'06, Proceeding of 2006 Workshop on Advanced Control and Diagnosis*.

Bekrar, R., Messai, N., Essounbouli, N., Hamzaoui, A., and Riera, B. (2006b). Off-line identification for a class of discrete event systems using safe Petri nets. In *DES-Des'06, 3rd IFAC Workshop on Discrete Event System Design*.

Boel, R., Ben-Naoum, L., and Breusegem, V. V. (1995). On forbidden state problems for a class of controlled Petri nets. In *IEEE Transaction on Automatic and Control*.

Ghaffari, A., Rezg, N., and Xie, X. (2002). Algebraic and geometric characterization of Petri net controllers using the theory of regions. In *WODES'02, Proceedings of 6$^{th}$ Worksop on Discrete Event Systems*.

Ghaffari, A., Rezg, N., and Xie, X. (2003). Feedback control logic for forbidden state problems of marked graphs: application to a real manufacturing system. In *IEEE transaction on Automatic and Control*.

Giua, A., DiCesare, F., and Silva, M. (1992). Generalized mutual exclusion constraints on nets with uncontrollable transitions. In *Proceedings of IEEE International Conference on System, Man and Cybernitic*.

Giua, A. and Xie, X. (2004). Control of safe ordinary Petri nets with marking specifications using unfolding. In *WODES'04, Proceedings of the 7$^{th}$ IFAC Workshop on Discrete Event Systems*.

Holloway, L., Guan, X., and Zhang, L. (1996). A generalization of state avoidance policies for controlled Petri nets. In *IEEE Transaction on Automatic and Control*.

Krogh, B. H. and Holloway, L. E. (1991). Synthesis of feedback control logic for discrete manufacturing systems. In *Automatica*.

Lee, E. J., Toguyeni, A., and Dangoumau, N. (2006). A Petri net based approach for the synthesis of parts' controllers for reconfigurable manufacturing systems. In *Proceeding of SICE-ICASE International Joint Conference*.

Moody, J. and Antsaklis, P. (2000). Petri net supervisors for dbs with uncontrollable and unobservable transitions. In *Proceedings of IEEE International Conference on System, Man and Cybernitic*.

# A HIGHER-ORDER STATISTICS-BASED VIRTUAL INSTRUMENT FOR TERMITE ACTIVITY TARGETING

Juan José González de la Rosa, José Melgar Camarero, Stephane Bouaud, J. G. Ramiro
*Univ. Cádiz, Electronics Area, Research Group PAI-TIC-168, EPSA, Av. Ramón Puyol S/N, E-11202-Algeciras-Cádiz, Spain*
*juanjose.delarosa@uca.es*

Antonio Moreno Muñoz
*Univ. Córdoba. Electronics Area. Research Group PAI-TIC-168*
*Campus Rabanales, A. Einstein C-2, E-14071, Córdoba, Spain*
*amoreno@uco.es*

Keywords: Acoustic Emission, Discrete Wavelet Transform, Higher-Order Statistics, Insect detection, Spectral kurtosis, Transient detection.

Abstract: In this paper we present the operation results of a portable computer-based measurement equipment conceived to perform non-destructive testing of suspicious termite infestations. Its signal processing module is based in the spectral kurtosis (SK), with the de-noising complement of the discrete wavelet transform (DWT). The SK pattern allows the targeting of alarms and activity signals. The DWT complements the SK, by keeping the successive approximations of the termite emissions, supposed more non-gaussian (less noisy) and with less entropy than the detail approximations. For a given mother wavelet, the maximum acceptable level, in the wavelet decomposition tree, which preserves the insects' emissions features, depends on the comparative evolution of the approximations details' entropies, and the value of the global spectral kurtosis associated to the approximation of the separated signals. The paper explains the detection criterion by showing different types of real-life recordings (alarms, activity, and background).

## 1 INTRODUCTION

Biological transients gather all the natural complexity of their associated sources, and the media through which they propagate. As a consequence, finding the most adequate method to get a complete characterization of the emission, implies the selection of the appropriate model, which better explains the processes of generation, propagation and capture of the emitted signals. This description matches the issue of measurement termite activity.

This paper deals with the performance of a final-version equipment (computer-based signal processing unit), whose previous prototype's performance, based in the time-frequency domain analysis of the kurtosis, was described in (De la Rosa and Muñoz, 2008, ). In this final version, the measurement method is mainly based in the interpretation of the spectral kurtosis graph, along with the wavelet analysis, which is thought as an aid. At the same time, we use a simple data acquisition unit, the sound card (maximum speed at 44,100 Hz), which simplifies the hardware unit and the criterion of detection.

The instruments for plague detection are thought with the objective of decreasing subjectiveness of the field operator. On-site monitoring implies reproducing the natural phenomenon of insect emissions with high accuracy. As a consequence it is imperative the use of a deep storage device, and high sensitive probes with selective frequency characteristics. These features make the price paid very high, and still do not guarantee the success of the detection. Besides, the expert's subjectiveness plays a crucial role, because only trained field operators can separate the signals of interest from the non-usable background.

Regarding the procedures, the methods in which the instruments are based are very much dependent on the detection of excess of power in the signals; these are the so-called second-order methods. For example, the RMS calculation can only characterize the intensity (amplitude level of the signal), and does not provide information regarding the envelope of the signal nor the time fluctuations of the amplitude. Another handicap of the second-order principle, e.g. the classical power spectrum, attends to the preservation of the energy during data processing. Consequently, the eradication of additive noise lies in filter design and sub-band decomposition, like wavelets and wavelet

packets.

As an alternative to improve noise rejection and complete characterization of the signals, in the past ten years, a myriad of higher-order methods are being applied in different fields of Science and Technology, in scenarios which involve signal separation and characterization of non-Gaussian signals. Concretely, the area of diagnostics-monitoring of rotating machines is also under our interest due to the similarities of the signals to be monitored with the transients from termites. Many time-series of faulty rotating machines consist of more-or-less repetitive short transients of random amplitudes and random occurrences of the impulses.

This paper describes a method based in the spectral kurtosis (related to the fourth-order cumulant at zero lags) to detect infestations of subterranean termites in a real-life scenario (southern Spain). Wavelet decomposition is used as an extra tool to aid detection from the preservation of the approximation of the signal, which is thought to be more Gaussian than the details.

The interpretation of the results is focussed on the classical peakedness of the statistical probability distribution associated to each frequency component of the signal, to get a measure of the distance from the Gaussian distribution. The spectral kurtosis serves as a twofold tool. First, it enhances non-Gaussian signals over the background. Secondly, it offers a more complete characterization of the transients emitted by the insects, providing the user with the probability associated to each frequency component.

The paper is structured as follows: in Section 2 a review on termite detection and relevant HOS experiences sets the foundations. In Section 3 we make a brief report on the definition of kurtosis; we use an unbiased estimator of the spectral kurtosis, successfully used in (De la Rosa and Muñoz, 2008, ), using a higher measurement bandwidth. Results are presented in Section 5. Finally, conclusions are drawn in Section 6.

# 2 TERMITE DETECTION AND HIGHER-ORDER STATISTICS

## 2.1 Subterranean Termites: Fundamentals

Termites have become a threat in all the modern countries, mainly due to the advent of central heating in the buildings. Cause more damage to homes in U.S.A. than storms and fire combined, on the average, there could be as many as 15 to 20 subterranean termite colonies per hectare, which means that for example a typical U.S.A. home may easily have three to four colonies situated under or around it. Colonies can contain up to 1,000,000 members (De la Rosa and Muñoz, 2008, ).

Termite detection has been gaining importance within the research community in the last two decades, mainly due to the urgent necessity of avoiding the use of harming termiticides, and to the joint use of new emerging techniques of detection and hormonal treatments (IGR[1] products), with the aim of performing an early treatment of the infestation. A localized partial infestation can be exterminated after two or three generations of the colony's members with the aid of these hormones, which stop chitin synthesis. A chitin synthesis inhibitor kills termites by inhibiting formation of a new exoskeleton when they shed their existing exoskeleton. As a direct consequence, the weakened unprotected *workers* stop feeding the *queen* termite of the colony, which dies of starvation, finishing the reproduction process, and consequently cutting any possible replacement of the members of the colony with a new generation. In this paper the specie *reticulitermes lucifugus* is under study.

## 2.2 Subterranean Termites: Detection Project towards HOS

The primary method of termite detection consists of looking for evidence of activity. But only about 25 percent of the building structure is accessible, and the conclusions depend very much on the level of expertise and the criteria of the inspector (De la Rosa and Muñoz, 2008, ),(Robbins et al., 1991). As a consequence, new techniques have been developed to remove subjectiveness and gain accessibility.

User-friendly equipment is being currently used in targeting subterranean insect infestations by means of temporal analysis of the vibratory data sequences [2]. An acoustic-emission (AE) sensor or an accelerometer is fixed to the suspicious structure. This class of instruments is based on the calculation of the root mean square (RMS) value of the vibratory waveform. The RMS value comprises information of the AE raw signal power during each time-interval of measurement (averaging time). This measurement strategy conveys a loss of potentially valuable information both in the time and in the frequency domain (De la Rosa and Muñoz, 2008, ).

[1] **I**nhibitor **G**rowth **R**egulators

[2] The system AED2000 (Acoustic Emission Consulting) has proven to be an advance in the detection of several insect species.

On the other hand, the use of the RMS value can be justified both by the difficulty of working with raw AE signals in the high-frequency range, and the scarce information about sources and propagation properties of the AE waves through the substratum. Noisy media and anisotropy makes even harder the implementation of new methods of calculation and measurement procedures. A more sophisticated family of instruments makes use of spectral analysis and digital filtering to detect and characterize vibratory signals (Mankin and Fisher, 2002).

Other complementary second-order tools, like wavelets and wavelet packets (time-dependent technique) concentrate on transients and non-stationary movements, making possible the detection of singularities and sharp transitions, by means of sub-band decomposition. The method has been proved under controlled laboratory conditions, up to a SNR=-30 dB (De la Rosa et al., 2006, ).

Higher-order statistics, are being widely used in several fields. The following are relevant due to the similarities of the problems they study. The spectral kurtosis has been successfully described and applied to the vibratory surveillance and diagnostics of rotating machines (Antoni, 2006a),(Antoni, 2006b), showing an inedit set of results that include kurtogram-based calculations of optimal band-pass filters and their performance in detecting two types of machinery faults (ball faults and outer race fault in rolling elements bearings); the kurtosis of the filtered signals is enhanced, which improves the detection of the fault type under study.

In the field of insect detection, the work presented in (De la Rosa and Muñoz, 2008, ) set the foundations of the present paper. The combined used of the spectral kurtosis and the time-domain sliding kurtosis showed marked features associated to termite emissions. In the frequency domain (sample frequency 64,000 Hz) three frequency zones were identified in the spectral kurtosis graph as evidence of infestation; two in the audio band (which will be also checked in the present paper) and one in the near ultrasound ($\approx$ 22 kHz). In the present paper the sample frequency was fixed to 44,100 Hz and the sound card was directly driven by MATLAB, which presents the results in an user-oriented interface, which is forwarded in Fig. 1. In the measurement situation shown in Fig. 1, the time-raw data contains alarms an activity signals from termites. This is a clear example of positive detection.

The developed virtual instrument also calculates and presents the spectrum (up-right graph) and the raw data (bottom-left). The field operator adds therefore visual information to the classical audio-based

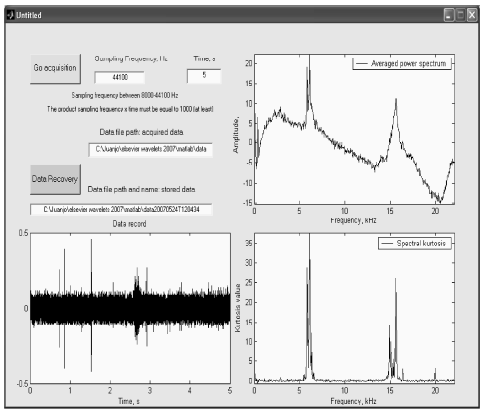



Figure 1: The graphical user interface which presents the results to the field operator. The spectral kurtosis is in the bottom-right corner.

criterion, which was by the way very subjective and very expertise-depend.

Other relevant achievements related to HOS are the following. Cumulants have been modeled in order to characterize the ultrasound waves in materials (Miralles et al., 2004). Bi-cepstrum, have been successfully used in blind identification of acoustic emissions (Iturrospe et al., 2005). Bi-spectrum has been applied to enhance reflections in ring-type samples of steel pipes, in a non-destructive testing frame (De la Rosa et al., 2007b, ).

In the field of termite detection, a cumulant-based independent component analysis algorithm has proven to separate termites' alarm signals from synthetics noise backgrounds (De la Rosa et al., 2005, ) in a blind source separation scenario. The information contained in the diagonal of the bi-spectrum data structure has proven to enhance the frequency pattern of the termites' emissions (De la Rosa et al., 2007a, ). The conclusions of these works were funded in the advantages of cumulants; in particular, in the capability of enhancing the SNR of a signal buried in noise processes, whose probability dnsity function is symmetrically distributed. The computational cost (memory consuming) could be pointed as the main drawback of the technique. Calculation of the cumulants' is made for all the combinations of time lags, giving rise to complex multidimensional data structures. The exam of this data sets leads to the selection of a privileged direction, whose which data are analyzed. In this paper, time-lags are set to zero in order to reduce the cost of computation. Statistically speaking, zero time lags lead to kurtosis, in a fourth-order cumulant.

# 3 KURTOSIS AND SPECTRAL KURTOSIS

## 3.1 Kurtosis, *4*th-order Cumulants and its Interpretation

Kurtosis is a measure of the "peakedness" of the probability distribution of a real-valued random variable. Higher kurtosis means more of the variance is due to infrequent extreme deviations, as opposed to frequent modestly-sized deviations. This fact is by the way used in this paper to detect termite emissions in an urban background. Kurtosis is more commonly defined as the fourth central cumulant divided by the square of the variance of the probability distribution, which is the so-called excess kurtosis:

$$\gamma_2 = \frac{\kappa_4}{\kappa_2^2} = \frac{\mu_4}{\sigma^4} - 3, \tag{1}$$

where $\mu_4 = \kappa_4 + 3\kappa_2^2$ is the *4*th-order central moment; and $\kappa_4$ is the *4*th-order central cumulant, i.d. the ideal value of $Cum_{4,x}(0,0,0)$. This definition of the *4*th-order cumulant for zero time-lags comes from a combinational relationship among the cumulants of stochastic signals and their moments, and is given by the *Leonov-Shiryaev* formula. A complete description for these statistics are found for example in (Nikias and Mendel, 1993; Mendel, 1991; Chonavel, 2003).

The "minus 3" at the end of this formula is a correction to make the kurtosis of the normal distribution equal to zero. Excess kurtosis can range from -2 to $+\infty$.

The sample kurtosis is calculated over a sample-register (an *N*-point data record), and noted by:

$$g_2 = \frac{m_4}{s^4} - 3 = \frac{m_4}{m_2^2} - 3 = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^4}{\frac{1}{N^2}\left[\sum_{i=1}^{N}(x_i - \bar{x})^2\right]^2} - 3, \tag{2}$$

where $m_4$ is the fourth sample moment about the mean, $m_2$ is the second sample moment about the mean (that is, the sample variance), and $\bar{x}$ is the sample mean. The sample kurtosis defined in Eq. (2) is a biased estimator of the population kurtosis, if we consider a sub-set of samples from the population (the observed data).

## 3.2 Spectral Kurtosis Estimation and Interpretation

Ideally, the spectral kurtosis is a representation of the kurtosis of each frequency component of a process (or data from a measurement instrument $x_i$). For estimation issues we will consider $M$ realizations of the process; each realization containing $N$ points; i.d. we consider $M$ measurement sweeps, each sweep with $N$ points. The time spacing between points is the sampling period, $T_s$, of the data acquisition unit.

A biased estimator for the spectral kurtosis for a number $M$ of *N*-point realizations at the frequency index $m$, is given by:

$$\hat{G}_{2,X}^{N,M}(m) = \\ = \frac{M}{M-1}\left[\frac{(M+1)\sum_{i=1}^{M}|X_N^i(m)|^4}{\left(\sum_{i=1}^{M}|X_N^i(m)|^2\right)^2} - 2\right]. \tag{3}$$

This estimator is the one we have implemented in the program code in order to perform the data computation and it was also used successfully in (Vrabie et al., 2003; De la Rosa and Muñoz, 2008, ).

Regarding the experimental signals, we expect to detect positive peaks in the kurtosis's spectrum, which may be associated to termite emissions, characterized by random-amplitude impulse-like events. This non-Gaussian behavior should be enhanced over the symmetrically distributed electronic noise, introduced in the measurement system. Speech is perhaps also reflected in the spectral kurtosis but not in the frequencies were termite emissions manifest. Besides, we assume, as a starting point, that non-Gaussian behavior of termite emissions is more acute than in speech. As a consequence, these emissions would be clearly outlined in the kurtosis spectrum. As a final remark, we expect that constant amplitude interferences are clearly differentiate due to their negative peaks in the spectral kurtosis. To show the ideal performance of the estimator, which has been described in these lines, and also described in (De la Rosa and Muñoz, 2008, ), we show an example based in synthetics. A mix of six different signals have been designed. Each mixture is the sum of a constant-amplitude sine of 2 kHz, a constant-amplitude sine at 9 kHz, a Gaussian-distributed-amplitude sine al 5 kHz, a Gaussian-distributed-amplitude sine at 18 kHz, a Gaussian white noise, and a colored Gaussian noise between 12 and 13 kHz. Each mixture (realization or sample register) contains 1324 points.

Negative kurtosis is expected for constant-amplitude processes, positive kurtosis should be associated to random-amplitudes and zero kurtosis will characterize both Gaussian-noise processes.

A simulation has been made in order to show the influence of the number of sample registers (M) in the averaged results for the SK graph. Fig. 2 shows a good performance because enough registers have

been averaged (M=500). For M ¡ 100, roughly, performance degenerates.

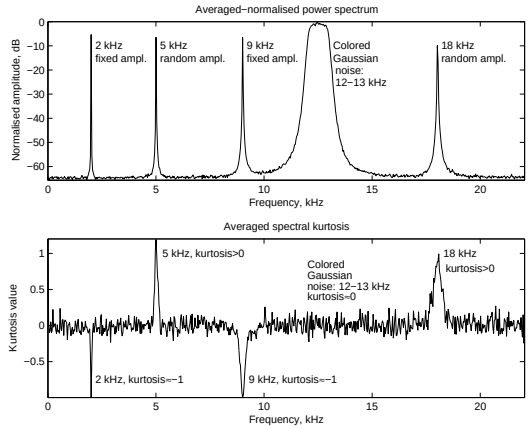



Figure 2: Performance over a set of synthetics, for M=500 realizations.

Once we have expose the main criterion used by the instrument, we make a brief recall of wavelets transforms, which are used as a complement in the following way. First of all, they manage to extract the impulses buried in noise and other parasitic signals. Secondly, the successive approximations of the signals, in the wavelet decomposition tree are supposed to gather all the non-Gaussian features or components, while the details are mainly composed by random signals, with a high entropy. So, the global SK (averaged SK, over all the frequency components) will be higher as the decomposition level increases.

# 4 THE WAVELET TRANSFORM DECOMPOSITION: FUNDAMENTALS AND DECISION CRITERIA

A *mother wavelet* is a function $\psi$ with finite energy[3], and zero average:

$$\int_{-\infty}^{+\infty} \psi(t)dt = 0, \tag{4}$$

This function is normalized[4], $\|\psi\| = 1$, and is centered in the neighborhood of t=0.

$\psi(t)$ can be expanded with a scale parameter $a$, and translated by $b$, resulting the *daughter functions* or *wavelet atoms*, which remain normalized:

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right); \tag{5}$$

The CWT can be considered as a correlation between the signal under study $s(t)$ and the wavelets (*daughters*). For a real signal $s(t)$, the definition of CWT is:

$$CWTs(a,b) = \frac{1}{\sqrt{a}}\int_{-\infty}^{+\infty} s(t)\psi^*\left(\frac{t-b}{a}\right)dt; \tag{6}$$

where $\psi^*(t)$ is the complex conjugate of the mother wavelet $\psi(t)$, $s(t)$ is the signal under study, and $a$ and $b$ are the scale and the position respectively ($a \in \Re^+ - 0, b \in \Re$). The scale parameter is proportional to the reciprocal of the frequency. Eq. (6) establishes that each coefficient provide numerical information about the similarity between the signal under study and the time-shifted frequency-scaled wavelet daughter.

The Discrete Time Wavelet Transform (DTWT) is introduced in order to to reduce the computational cost of calculating all these coefficients. Only a subset of scale and time shifts are chosen in the DTWT. A tree-structure arrangement of filters allows the subband decomposition of the signal. The original signal passes through two complementary filters (*quadrature mirror filters*), and two signals are obtained as a result of a down-sampling process, corresponding to the approximation and detail coefficients.

The lengths of the detail and approximation coefficient vectors are slightly more than half the length of the original signal, $s(t)$. This is the result of the digital filtering process (convolution) (Angrisani et al., 1999). The approximations are the high-scale, low-frequency components of the signal. The details are the low-scale, high-frequency components.

*Daubechies* 5 has been selected as most similar wavelet mother, because of the highest coefficients in the decomposition tree. Given the wavelet mother, to show the process of selecting the maximum decomposition level in the wavelet tree, we have adopted a criterion based on the calculation of Shannon's entropy (information entropy), which is a measure of the uncertainty associated with a random variable $X$; this entropy denoted by $H(X)$, and defined by:

$$H(X) := -\sum_{i=1}^{N} p(x_i)\log_{10} p(x_i), \tag{7}$$

where $X$ is an N-outcome measurement process $\{x_i, i = 1, \cdots, N\}$, and $p(x_i)$ is the probability density function of the outcome $x_i$.

We show this strategy via the following example, based on real-life data, presented in Fig. 6 and in Fig.

[3] $f \in \mathbf{L}^2(\Re)$, the space of the finite energy functions, verifying $\int_{-\infty}^{+\infty} |f(t)|^2\, dt < +\infty$.

[4] $\|f\| = \left(\int_{-\infty}^{+\infty} |f(t)|^2\, dt\right)^{1/2} = 1$.

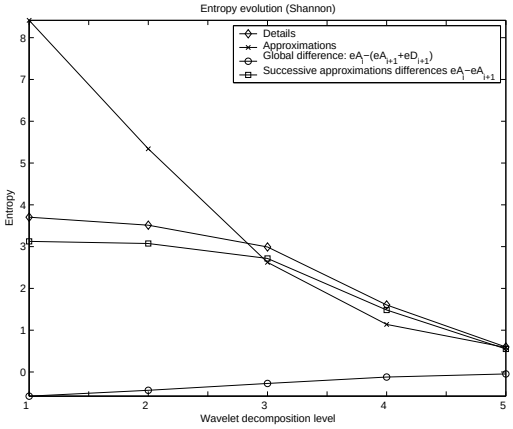



Figure 3: Evolution of the entropy.

7, in the results section. The entropy of the approximations and the details are compared for each level of comparison and shown in Fig. 3.

By looking at the graph of Fig. 3, at level 4, the entropy of the approximations is less than the entropy of the details. So level 4 is in a sense, a point of inversion. No improvement is obtained for level 5, where the entropies are very similar.

We can also see that the global difference of entropies increases towards zero, at level 5, as a complementary indication that further decomposition will not suppose progress in de-noising.

# 5 EXPERIMENTS AND RESULTS

## 5.1 The Instrument and the Measurement Procedure

A piezoelectric probe-sensor (model SP-1L from *Acoustic Emission Consulting*) is used in the final version of the instrument, and was described in detail in (De la Rosa and Muñoz, 2008, ). The sensor is connected to the sound card of a lap-top computer and the acquisition is driven by MATLAB, via the Graphical User Interface (GUI).

The user interface was presented in Fig. 1. The operator can select the acquisition time and the sample frequency (maximum 44,100 Hz if the sound card is driven). In the bottom-right corner of Fig. 1, the spectral kurtosis graph is presented. The user can also examine the raw data and the spectrum. Automatically, the instrument save the acquired data (labeling the file with the date). Additionally, the operator can recall the stored files.

The transducer SP-1L was used to record the data registers in the field experience, and the ICP unit (**I**ntegrated **C**ircuit **P**iezoelectric; ICP interface) was connected to the sound card of a lap-top computer, configuring an autonomous measurement unit. The sampling frequency was $F_s$=44,100 Hz for all the registers analyzed in this paper, both in the sliding-cumulant results as in the spectral kurtosis subsection. The recording stage took place in a garden with evidence of infestation and the bare waveguide of the sensor was introduced in the lawn, over the suspicious zone. Termite sounds from feeding are like sharp pops and crackles in the audio output.

The key of the spectral kurtosis detection strategy used in this work lies in the potential enhancement of the non-Gaussian behavior of the emissions. If this happens, i.e. if an increase of the non-Gaussian activity (increase in the kurtosis, peakedness of the probability distribution) is observed-measured in the spectral kurtosis graph, there may be infestation in the surrounding subterranean perimeter, where the transducer is attached.

Termite emissions are non-stationary, so the instrument treats data by ensemble averaging of the sample registers, following the indications in (Bendat and Piersol, 2000) (pp. 463-465). Each spectrum and spectral kurtosis graph presented in this section is the result of averaging the spectra of the sample registers, or realizations. As a final remark, acquired data is normalized according to the norm: $\|s\| = \left(\sum_{i=1}^{N} |s_i|^2\right)^{1/2}$.

## 5.2 Operating Cases

In this subsection we present the possible situations associated to the measurement cases. We present the signals out of the instrument display in order to be analyzed more precisely. A data acquisition time of 5 seconds and a sample frequency of 44,100 Hz have been selected. So every time the user performs an acquisition (pressing the button ”Go”) 220,500 points are stored. The software-engine is adjusted to calculate the averaged spectral kurtosis (SK) over a set of 220 realizations, each of them containing 1,000 points.

Two couples of data registers have been selected as significant examples, corresponding to typical measurements situations. For a given couple, first we present the results without applying wavelets. Then we explain the information wavelets add.

Fig. 4 presents a clear detection case, characterized by termite activity signals without alarms. Two peaks are clearly enhanced in the SK graph (near 5 kHz, and near 15 kHz).

The de-noised data in the time main are shown in the upper grpah of Fig. 5. Applying the spec-

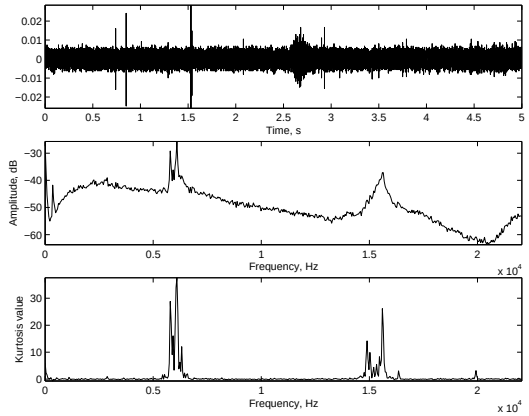



Figure 4: A clear measurement of activity detection.

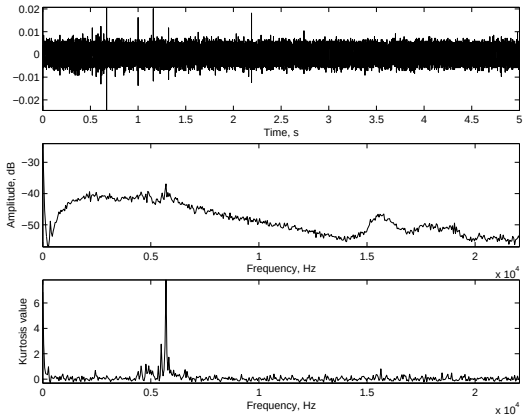



Figure 6: A doubtful measurement situation.

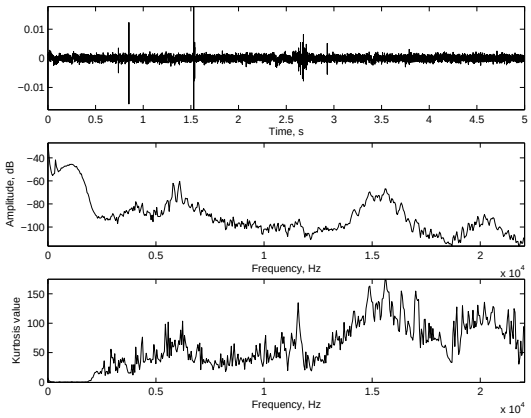



Figure 5: De-noising results for data in Fig. 4. A general enhancement of the spectral kurtosis occurs.

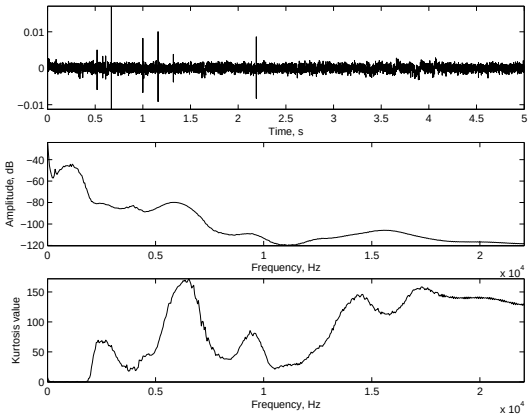



Figure 7: De-noising results of data in Fig. 6.

tral kurtosis to the de-noised version it is seen that all the frequency components are enhanced, specially those ones in the detection band. This fact confirms the presence of insects, and it is of special value in doubtful situations, when they are really needed.

In Fig. 6 a doubtful measurement case is presented. Activity evidence is outlined only near 5 kHz. Once, the wavelets have been applied (shown in Fig. 7), the enhancement near 5 kHz and 15 kHz confirm the detection.

Hereinafter, we present the conclusions.

# 6 CONCLUSIONS AND ACCOMPLISHMENTS

Assuming the starting hypothesis that the insect emissions may have a more peaked probability distribution than any other simultaneous source of emission in the measurement perimeter, we have design a termite detection strategy and a virtual instrument based in the calculation of the *4*th-order cumulants for zero time lags, which are indicative of the signals' kurtosis. The instrument is actually in use by an Spanish company.

An estimator of the spectral kurtosis has been used to perform a selective analysis of the peakedness of the signal. It has been shown that new frequency components gain in relevance in the spectral kurtosis graphs.

The main goal of this signal-processing method is to reduce subjectiveness due to visual or listening inspection of the registers. This means that in a noisy environment, it may be possible to ignore termite feeding activity even with an *ad hoc* sensor because, despite the fact that the sensor is capable of register these low-level emissions, the human ear can easily ignore them (De la Rosa and Muñoz, 2008, ).

## ACKNOWLEDGEMENTS

The authors would like to thank the *Spanish Ministry of Science and Education* for funding the project DPI2003-00878, where the different noise processes have been modeled and contrasted; and also for supporting the PETRI project PTR95-0824-OP dealing with plague detection using higher-order statistics. Our unforgettable thanks to the trust we have from the *Andalusian Government* for funding the excellency project PAI2005-TIC00155, where higher-order statistics are modeled and applied to plague detection and power quality analysis.

## REFERENCES

Angrisani, L., Daponte, P., and D'Apuzzo, M. (1999). A method for the automatic detection and measurement of transients. part I: the measurement method. *Measurement*, 25(1):19–30.

Antoni, J. (2006a). The spectral kurtosis: a useful tool for characterising non-stationary signals. *Mechanical Systems and Signal Processing (Ed. Elsevier)*, 20(2):282–307.

Antoni, J. (2006b). The spectral kurtosis: application to the vibratory surveillance and diagnostics of rotating machines. *Mechanical Systems and Signal Processing (Ed. Elsevier)*, 20(2):308–331.

Bendat, J. and Piersol, A. (2000). *Random Data Analysis and Measurement Procedures*, volume 1 of *Wiley Series in Probability and Statistics*. Wiley Interscience, 3 edition.

Chonavel, T. (2003). *Statistical Signal Processing. Modelling and Estimation*, volume 1 of *Advanced Textbooks in Control and Signal Processing*. Springer, London, 1 edition.

Iturrospe, A., Dornfeld, D., Atxa, V., and Abete, J. M. (2005). Bicepstrum based blind identification of the acoustic emission (AE) signal in precision turning. *Mechanical Systems and Signal Processing (Ed. Elsevier)*, 19(1):447–466.

Mankin, R. W. and Fisher, J. R. (2002). Current and potential uses of acoustic systems for detection of soil insects infestations. In *Proceedings of the Fourth Symposium on Agroacoustic*, pages 152–158.

Mendel, J. M. (1991). Tutorial on higher-order statistics (spectra) in signal processing and system theory: Theoretical results and some applications. *Proceedings of the IEEE*, 79(3):278–305.

Miralles, R., Vergara, L., and Gosalbez, J. (2004). Material grain noise analysis by using higher-order statistics. *Signal Processing (Ed. Elsevier)*, 84(1):197–205.

Nikias, C. L. and Mendel, J. M. (1993). Signal processing with higher-order spectra. *IEEE Signal Processing Magazine*, pages 10–37.

Robbins, W. P., Mueller, R. K., Schaal, T., and Ebeling, T. (1991). Characteristics of acoustic emission signals generated by termite activity in wood. In *Proceedings of the IEEE Ultrasonic Symposium*, pages 1047–1051.

De la Rosa, J. J. G., Lloret, I., Moreno, A., Puntonet, C. G., and Górriz, J. M. (2006). Wavelets and wavelet packets applied to detect and characterize transient alarm signals from termites. *Measurement (Ed. Elsevier)*, 39(6):553–564. Available online 10 January 2006.

De la Rosa, J. J. G., Lloret, I., Puntonet, C. G., Piotrkowski, R., and Moreno, A. (2007a). Higher-order spectra measurement techniques of termite emissions. a characterization framework. *Measurement (Ed. Elsevier)*, In Press:–. Available online 13 October 2006.

De la Rosa, J. J. G. and Muñoz, A. M. (2008). Higher-order cumulants and spectral kurtosis for early detection of subterranean termites. *Mechanical Systems and Signal Processing (Ed. Elsevier)*, 22(Issue 1):279–294. Available online 1 September 2007.

De la Rosa, J. J. G., Piotrkowski, R., and Ruzzante, J. (2007b). Third-order spectral characterization of acoustic emission signals in ring-type samples from steel pipes for the oil industry. *Mechanical Systems and Signal Processing (Ed. Elsevier)*, 21(Issue 4):1917–1926. Available online 10 October 2006.

De la Rosa, J. J. G., Puntonet, C. G., and Lloret, I. (2005). An application of the independent component analysis to monitor acoustic emission signals generated by termite activity in wood. *Measurement (Ed. Elsevier)*, 37(1):63–76. Available online 12 October 2004.

Vrabie, V., Granjon, P., and Serviere, C. (2003). Spectral kurtosis: from definition to application. In IEEE, editor, *IEEE-EURASIP International Workshop on Nonlinear Signal and Image Processing (NSIP'2003)*, volume 1, pages 1–5.

# A RECURSIVE FRISCH SCHEME ALGORITHM FOR COLOURED OUTPUT NOISE

J. G. Linden and K. J. Burnham
*Control Theory and Applications Centre, Coventry University, Priory Street, Coventry, U.K.*
*j.linden@coventry.ac.uk*

Keywords: Adaptive Estimation, Errors-in-varibles, Recursive Identification, System Identification.

Abstract: A recursive (adaptive) algorithm for the identification of dynamical linear errors-in-variables systems in the case of coloured output noise is developed. The input measurement noise variance as well as the auto-covariance elements of the coloured output noise sequence are determined via two separate Newton algorithms. The model parameter estimates are obtained by a recursive bias-compensating instrumental variables algorithm with past noisy inputs as instruments, thus allowing the compensation for the explicitly computed bias at each discrete-time instance. The performance of the developed algorithm is demonstrated via simulation.

## 1 INTRODUCTION

Linear time-invariant (LTI) errors-in-variables (EIV) models are characterised by an exact linear relationship between input and output signals where both quantities 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), which yields estimates of the model parameters as well as the measurement noise variances. The dynamic Frisch scheme presented in (Beghelli et al., 1990; Söderström, 2007a) assumes that the additive disturbances on the system input and output are white. Such an assumption, however, can be rather restrictive since the output noise often not solely consists of measurement uncertainties, but also aims to account for process disturbances, which are usually correlated in time. In order to overcome this shortcoming, the Frisch scheme has recently been extended to the coloured output noise case (Söderström, 2008). This paper develops a recursive (adaptive) formulation of the algorithm developed in (Söderström, 2008), which allows the estimates to be calculated online as new data arrives. Recursive algorithms for the white noise case have been considered in (Linden et al., 2008; Linden et al., 2007).

The paper is organised as follows. Section 2 presents the EIV identification problem and introduces some notational conventions. Section 3 reviews the offline Frisch scheme procedure for the white noise as well as the coloured noise case. Section 4 develops the recursive algorithm and Section 5 provides a numerical example. Conclusions are given in Section 6.

## 2 PROBLEM STATEMENT AND NOTATION

In this paper, a discrete-time, LTI single-input single-output (SISO) EIV system is considered, which is described by

$$A(q^{-1})y_{0_i} = B(q^{-1})u_{0_i}, \tag{1}$$

where $i$ is an integer valued time index and

$$A(q^{-1}) \triangleq 1 + a_1 q^{-1} + \ldots + a_{n_a} q^{-n_a}, \tag{2a}$$
$$B(q^{-1}) \triangleq b_1 q^{-1} + \ldots + b_{n_b} q^{-n_b} \tag{2b}$$

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

$$u_i = u_{0_i} + \tilde{u}_i, \tag{3a}$$
$$y_i = y_{0_i} + \tilde{y}_i \tag{3b}$$

are available, where $\tilde{u}_i$ and $\tilde{y}_i$ denote input and output measurement noise, respectively. Such a 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.

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.

A5a. The sequence $\tilde{u}_i$ is a zero-mean, ergodic, white noise process with unknown variance $\sigma_{\tilde{u}}$.

A5b. The sequence $\tilde{y}_i$ is a zero-mean, ergodic noise process with unknown auto-covariance sequence $\{r_{\tilde{y}}(0), r_{\tilde{y}}(1), \cdots\}$.

A6. The noise sequences $\tilde{u}_i$ and $\tilde{y}_i$ are mutually uncorrelated and uncorrelated with $u_{0_i}$.

The auto-covariance elements in A5b are defined by

$$r_{\tilde{y}}(\tau) \triangleq E\left[\tilde{y}_k \tilde{y}_{k-\tau}\right], \tag{4}$$

where $E[\cdot]$ denotes the expected value operator. Within this paper covariance matrices of two column vectors $v_k$ and $w_k$ are denoted

$$\Sigma_{vw} \triangleq E\left[v_k w_k^T\right], \qquad \Sigma_v \triangleq E\left[v_k v_k^T\right], \tag{5}$$

whilst vectors consisting of covariance elements are denoted

$$\xi_{vc} \triangleq E\left[v_k c_k\right] \tag{6}$$

with $c_k$ being a scalar. The corresponding estimated sample covariance elements are denoted in a similar manner

$$\hat{\Sigma}_{vw}^k \triangleq \frac{1}{k}\sum_{i=1}^{k} v_k w_k^T, \quad \hat{\Sigma}_v^k \triangleq \frac{1}{k}\sum_{i=1}^{k} v_k v_k^T, \quad \hat{\xi}_{vc}^k \triangleq \frac{1}{k}\sum_{i=1}^{k} v_k c_k. \tag{7}$$

In addition, the parameter vectors are formed by

$$\theta \triangleq \begin{bmatrix} a^T & b^T \end{bmatrix}^T = \begin{bmatrix} a_1 & \ldots & a_{n_a} & b_1 & \ldots & b_{n_b} \end{bmatrix}^T, \tag{8a}$$

$$\bar{\theta} \triangleq \begin{bmatrix} \bar{a}^T & b^T \end{bmatrix}^T = \begin{bmatrix} 1 & \theta^T \end{bmatrix}^T, \tag{8b}$$

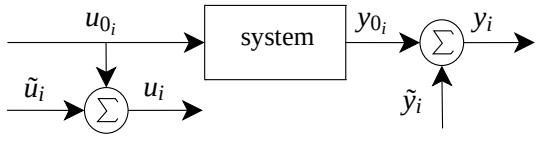



Figure 1: Errors-in-variables setup.

which gives an alternative description of (1)-(3) by

$$\bar{\varphi}_{0_i}^T \bar{\theta} = 0, \tag{9a}$$

$$\bar{\varphi}_i = \bar{\varphi}_{0_i} + \tilde{\bar{\varphi}}_i, \tag{9b}$$

where the regression vector is given by

$$\varphi_i \triangleq \begin{bmatrix} \varphi_{y_i}^T & \varphi_{u_i}^T \end{bmatrix}^T \tag{10}$$

$$\triangleq \begin{bmatrix} -y_{i-1} & \ldots & -y_{i-n_a} & u_{i-1} & \ldots & u_{i-n_b} \end{bmatrix}^T,$$

$$\bar{\varphi}_i \triangleq \begin{bmatrix} \bar{\varphi}_{y_i}^T & \varphi_{u_i}^T \end{bmatrix}^T \triangleq \begin{bmatrix} -y_i & \varphi_i^T \end{bmatrix}^T. \tag{11}$$

The noise-free regression vectors $\varphi_{0_i}$, $\bar{\varphi}_{0_i}$ and the vectors containing the noise contributions $\tilde{\varphi}_i$ , $\tilde{\bar{\varphi}}_i$ are defined in a similar manner. The identification problem is now given by:

**Problem 1.** *Given k samples of noisy input-output data* $\{u_1, y_1, \ldots, u_k, y_k\}$*, determine an estimate of the augmented parameter vector*

$$\vartheta \triangleq \begin{bmatrix} a_1 & \ldots & a_{n_a} & b_1 & \ldots & b_{n_b} & \sigma_{\tilde{u}} \\ & & r_{\tilde{y}}(0) & \cdots & r_{\tilde{y}}(n_a) & & \end{bmatrix}^T. \tag{12}$$

Throughout this paper, the convention is made that estimated quantities are marked by a $\hat{}$ whilst time dependent quantities have a sub- or superscript $k$, e.g. $\hat{\Sigma}_{\varphi}^k$ for a sample covariance matrix corresponding to $\Sigma_{\varphi}$.

# 3 FRISCH SCHEME

## 3.1 White Noise Case

If the least squares (LS) estimator is directly applied to estimate the parameters of the EIV system (1)-(3), the estimates will generally be biased (Söderström, 2007a). However, if the statistical nature of the noise sequences is known, it is possible to compensate for the bias. The Frisch scheme belongs to the class of such bias-compensating LS techniques. The compensated normal equations are given by

$$\left(\hat{\Sigma}_{\varphi}^k - \hat{\Sigma}_{\tilde{\varphi}}^k\right)\hat{\theta}_k = \hat{\xi}_{\varphi y}^k, \tag{13}$$

where $\hat{\Sigma}_{\varphi}^k$ and $\hat{\xi}_{\varphi y}^k$ are defined by (7). In the case of white noise sequences the compensating matrix $\hat{\Sigma}_{\tilde{\varphi}}^k$ is diagonal and given by

$$\begin{bmatrix} \hat{r}_{\tilde{y}}^k(0) I_{n_a} & 0 \\ 0 & \hat{\sigma}_{\tilde{u}}^k I_{n_b} \end{bmatrix}, \tag{14}$$

where $I_n$ denotes the identity matrix of dimension $n$. Within the Frisch scheme, the variances $\hat{\sigma}_{\tilde{u}}^k$, $\hat{r}_{\tilde{y}}^k(0)$ of input and output measurement noise, respectively, are

determined such that the extended compensated normal equations equate to zero

$$0 = \left(\hat{\Sigma}^k_{\bar{\varphi}} - \hat{\Sigma}^k_{\tilde{\varphi}}\right)\hat{\tilde{\theta}}_k \tag{15}$$
$$= \left(\begin{bmatrix} \hat{\Sigma}^k_{\varphi_y} & \hat{\Sigma}^k_{\varphi_y\varphi_u} \\ \hat{\Sigma}^k_{\varphi_u\bar{\varphi}_y} & \hat{\Sigma}^k_{\varphi_u} \end{bmatrix} - \begin{bmatrix} \hat{r}^k_{\tilde{y}}(0) I_{n_a+1} & 0 \\ 0 & \hat{\sigma}^k_{\tilde{u}} I_{n_b} \end{bmatrix}\right)\hat{\tilde{\theta}}_k,$$

i.e. such that $\hat{\Sigma}^k_{\bar{\varphi}} - \hat{\Sigma}^k_{\tilde{\varphi}}$ is singular. By utilising the Schur complement, the input noise variance can be expressed as a nonlinear function of the output noise variance and vice versa (Beghelli et al., 1990)

$$\hat{r}^k_{\tilde{y}}(0) = \lambda_{\min}\left(\hat{\Sigma}^k_{\varphi_y} - \hat{\Sigma}^k_{\varphi_y\varphi_u}\left[\hat{\Sigma}^k_{\varphi_u} - \sigma_{\tilde{u}} I_{n_b}\right]^{-1}\hat{\Sigma}^k_{\varphi_u\bar{\varphi}_y}\right), \tag{16a}$$

$$\hat{\sigma}^k_{\tilde{u}} = \lambda_{\min}\left(\hat{\Sigma}^k_{\varphi_u} - \hat{\Sigma}^k_{\varphi_u\bar{\varphi}_y}\left[\hat{\Sigma}^k_{\varphi_y} - r_{\tilde{y}}(0) I_{n_a+1}\right]^{-1}\hat{\Sigma}^k_{\varphi_y\varphi_u}\right), \tag{16b}$$

where $\lambda_{\min}$ denotes the minimum eigenvalue operator. Equation (16) together with (15) defines a whole set of possible solutions depending on the choice of $\sigma_{\tilde{u}}$ or $r_{\tilde{y}}(0)$, respectively. In order to uniquely solve the identification problem, another equation is required. Several choices are discussed in (Hong et al., 2007).

## 3.2 Coloured Noise Case

Now assume that $\tilde{y}_k$ is no longer white, i.e. it is correlated or coloured. For this case, the matrices $\hat{\Sigma}^k_{\bar{\varphi}}$ and $\hat{\Sigma}^k_{\tilde{\varphi}}$ in (15) can be expressed in block form as

$$\hat{\Sigma}^k_{\bar{\varphi}} = \begin{bmatrix} \times & \times & \times \\ -\hat{\xi}^k_{\varphi_y y} & \hat{\Sigma}^k_{\varphi_y} & \hat{\Sigma}^k_{\varphi_y\varphi_u} \\ -\hat{\xi}^k_{\varphi_u y} & \hat{\Sigma}^k_{\varphi_u\varphi_y} & \hat{\Sigma}^k_{\varphi_u} \end{bmatrix}, \tag{17a}$$

$$\hat{\Sigma}^k_{\tilde{\varphi}} = \begin{bmatrix} \times & \times & \times \\ -\hat{\xi}^k_{\tilde{\varphi}_y\tilde{y}} & \hat{\Sigma}^k_{\tilde{\varphi}_y} & 0 \\ 0 & 0 & \hat{\sigma}_{\tilde{u}} I^k_{n_b} \end{bmatrix}, \tag{17b}$$

where the first row consists of arbitrary entries $\times$ and

$$\hat{\Sigma}^k_{\tilde{\varphi}_y} = \begin{bmatrix} \hat{r}^k_{\tilde{y}}(0) & \cdots & \hat{r}^k_{\tilde{y}}(n_a-1) \\ \vdots & \ddots & \vdots \\ \hat{r}^k_{\tilde{y}}(n_a-1) & \cdots & \hat{r}^k_{\tilde{y}}(0) \end{bmatrix} \tag{18}$$

is a dense matrix, whilst the remaining entries in (17) are in accordance with (7). Consequently, the $n_a + n_b$ compensated normal equations in the case of correlated output noise are given by

$$\left(\begin{bmatrix} \hat{\Sigma}^k_{\varphi_y} & \hat{\Sigma}^k_{\varphi_y\varphi_u} \\ \hat{\Sigma}^k_{\varphi_u\varphi_y} & \hat{\Sigma}^k_{\varphi_u} \end{bmatrix} - \begin{bmatrix} \hat{\Sigma}^k_{\tilde{\varphi}_y} & 0 \\ 0 & \hat{\sigma}^k_{\tilde{u}} I_{n_b} \end{bmatrix}\right)\hat{\theta}_k = \begin{bmatrix} \hat{\xi}^k_{\varphi_y y} - \hat{\xi}^k_{\tilde{\varphi}_y\tilde{y}} \\ \hat{\xi}^k_{\varphi_u y} \end{bmatrix}. \tag{19}$$

Now consider the Frisch equation (16b) which becomes

$$\hat{\sigma}^k_{\tilde{u}} = \lambda_{\min}(B_k), \tag{20}$$

with

$$B_k \triangleq \hat{\Sigma}^k_{\varphi_u} - \hat{\Sigma}^k_{\varphi_u\bar{\varphi}_y}\left[\hat{\Sigma}^k_{\varphi_y} - \hat{\Sigma}^k_{\tilde{\varphi}_y}\right]^{-1}\hat{\Sigma}^k_{\varphi_y\varphi_u} \tag{21}$$

and it remains to specify $n_a + 1$ equations for the determination the auto-covariance elements

$$\hat{\rho}^k_y \triangleq \begin{bmatrix} \hat{r}^k_{\tilde{y}}(0) & \hat{r}^k_{\tilde{y}}(1) & \cdots & \hat{r}^k_{\tilde{y}}(n_a) \end{bmatrix}^T. \tag{22}$$

In (Söderström, 2008) several possibilities to obtain the remaining equations are discussed. It is shown that a covariance-matching criterion, as used in (Diversi et al., 2003), as well as correlating the residuals with past outputs, which corresponds to an instrumental variable (IV) -like approach with outputs as instruments, cannot be successful since it always leads to more unknowns than equations. However, by correlating the residuals, denoted $\varepsilon_k$, with past inputs, the remaining equations are obtained for the asymptotic case via

$$E\left[\bar{\zeta}_k \varepsilon_k\right] = 0, \tag{23}$$

where the instruments are given by

$$\bar{\zeta}_k = \begin{bmatrix} u_{k-n_b-1} & \cdots & u_{k-n_b-l} \end{bmatrix}^T \tag{24}$$

and the residuals are obtained via

$$\varepsilon_k = A(q^{-1})y_k - B(q^{-1})u_k = y_k - \varphi_k^T\theta. \tag{25}$$

This yields

$$\xi_{\bar{\zeta}y} - \Sigma_{\bar{\zeta}\varphi}\theta = 0, \tag{26}$$

which can be expressed in block form, and using sample statistics, as

$$\begin{bmatrix} \hat{\Sigma}^k_{\zeta\varphi_y} & \hat{\Sigma}^k_{\zeta\varphi_u} \end{bmatrix}\hat{\theta}_k = \hat{\xi}^k_{\bar{\zeta}y}, \tag{27}$$

where the length $l$ of the instrument vector $\bar{\zeta}_k$ must satisfy $l \geq n_a + 1$ in order to obtain at least $n_a + 1$ additional equations for the determination of $\hat{\rho}^k_y$.

In (Söderström, 2008), two algorithms have been proposed to solve the resulting (nonlinear) estimation problem. Here, the two step algorithm, which makes use of the separable LS technique is considered. Whilst in the white noise case the estimate of $\theta$ is obtained from the compensated normal equations after the noise variances have been determined, this approach is conceptually different as outlined in the remainder of this Section.

#### 3.2.1 Step 1

Note that $\hat{\rho}_y^k$ only appears in the first $n_a$ equations of (19) and by combining the last $n_b$ equations of the LS normal equations (19) with the $n_a+1$ IV equations (27), one can express

$$\begin{bmatrix} \hat{\Sigma}^k_{\varphi_u\varphi_y} & \hat{\Sigma}^k_{\varphi_u} - \hat{\sigma}^k_{\tilde{u}} I_{n_b} \\ \hat{\Sigma}^k_{\zeta\varphi_y} & \hat{\Sigma}^k_{\zeta\varphi_u} \end{bmatrix} \hat{\theta}_k = \begin{bmatrix} \hat{\xi}^k_{\varphi_u y} \\ \hat{\xi}^k_{\zeta y} \end{bmatrix}. \quad (28)$$

which constitute $n_a+n_b+1$ equations in $n_a+n_b+1$ unknowns ($\hat{\theta}$ and $\hat{\sigma}^k_{\tilde{u}}$). Equation (28) is an overdetermined system of normal equations with its first part obtained from the bias compensated LS and the second part given by the IV estimator, which uses delayed inputs. Moreover, it is nonlinear due to the multiplication of $\hat{\theta}_k$ with $\hat{\sigma}^k_{\tilde{u}}$.

In order to estimate $\theta$ and $\sigma_{\tilde{u}}$, (28) can be re-expressed as

$$\left(\hat{\Sigma}^k_{\bar{\delta}\varphi} - \hat{\sigma}^k_{\tilde{u}} \bar{J}\right) \hat{\theta}_k = \hat{\xi}^k_{\bar{\delta}y}, \quad (29)$$

where $\hat{\Sigma}^k_{\bar{\delta}\varphi}$ and $\hat{\xi}^k_{\bar{\delta}y}$ are defined by (7) with

$$\begin{aligned} \bar{\delta}_i &\triangleq \begin{bmatrix} \varphi_{u_i}^T & \bar{\zeta}_i^T \end{bmatrix}^T \\ &= \begin{bmatrix} u_{i-1} & \dots & u_{i-n_b} & u_{i-n_b-1} & \dots & u_{i-n_b-l} \end{bmatrix}^T, \end{aligned} \quad (30)$$

whilst $\bar{J}$ is given by

$$\bar{J} \triangleq \begin{bmatrix} 0 & I_{n_b} \\ 0 & 0 \end{bmatrix}. \quad (31)$$

Note that (29) can be interpreted as a bias-compensated IV approach, where the instrument vector $\bar{\delta}_i$ is constructed from past measured inputs. Introducing for convenience

$$\hat{G}_k \triangleq \hat{\Sigma}^k_{\bar{\delta}\varphi} - \hat{\sigma}^k_{\tilde{u}} \bar{J}, \quad (32)$$

the estimates for $\sigma^k_{\tilde{u}}$ and $\theta_k$ are obtained by minimising the (nonlinear) LS costfunction

$$\min_{\hat{\theta}_k, \hat{\sigma}^k_{\tilde{u}}} \left|\left| \hat{G}_k \hat{\theta}_k - \hat{\xi}^k_{\bar{\delta}y} \right|\right|^2 \quad (33)$$

which is minimised w.r.t. $\sigma^k_{\tilde{u}}$ and $\theta_k$. If $\hat{\sigma}^k_{\tilde{u}}$ is assumed to be fixed, an explicit expression for $\hat{\theta}_k$ is given by the well-known LS solution

$$\hat{\theta}_k = \hat{G}_k^{\dagger} \hat{\xi}^k_{\bar{\delta}y}, \quad (34)$$

where $\hat{G}_k^{\dagger} \triangleq (\hat{G}_k^T \hat{G}_k)^{-1} \hat{G}_k^T$ denotes the Moore-Penrose pseudo inverse. Using the separable LS approach (Ljung, 1999, p. 335), the problem is reduced to an optimisation in one variable only by substituting (34) in (33). Consequently, $\hat{\sigma}^k_{\tilde{u}}$ can be obtained via

$$\hat{\sigma}^k_{\tilde{u}} = \arg\min_{\hat{\sigma}^k_{\tilde{u}}} V_k \quad (35)$$

with

$$\begin{aligned} V_k &= \left|\left| \hat{G}_k \hat{G}_k^{\dagger} \hat{\xi}^k_{\bar{\delta}y} - \hat{\xi}^k_{\bar{\delta}y} \right|\right|^2 \\ &= \left[\hat{\xi}^k_{\bar{\delta}y}\right]^T \hat{\xi}^k_{\bar{\delta}y} - \left[\hat{\xi}^k_{\bar{\delta}y}\right]^T \hat{G}_k \left[\hat{G}_k^T \hat{G}_k\right]^{-1} \hat{G}_k^T \hat{\xi}^k_{\bar{\delta}y}. \end{aligned} \quad (36)$$

Once $\hat{\sigma}^k_{\tilde{u}}$ is obtained, $\hat{\theta}_k$ is given by (34). Since the solution of (35) should satisfy $V_k = 0$, the value of $V_k$ indicates whether the optimisation algorithm has converged to a global or local minimum (Söderström, 2008).

#### 3.2.2 Step 2

In order to determine the estimates for the autocorrelation sequence $\hat{\rho}_y^k$ the remaining $n_a$ normal equations

$$\left[\hat{\Sigma}^k_{\varphi_y} - \hat{\Sigma}^k_{\tilde{\varphi}_y} \quad \hat{\Sigma}^k_{\varphi_y\varphi_u}\right] \hat{\theta}_k = \hat{\xi}^k_{\varphi_y y} - \hat{\xi}^k_{\tilde{\varphi}_y \tilde{y}} \quad (37)$$

together with the Frisch equation (20) are considered. Equation (37) can be expressed as

$$\hat{\Sigma}^k_{\tilde{\varphi}_y} \hat{a}_k - \hat{\xi}^k_{\tilde{\varphi}_y \tilde{y}} = \left[\hat{\Sigma}^k_{\varphi_y} \quad \hat{\Sigma}^k_{\varphi_y\varphi_u}\right] \hat{\theta}_k - \hat{\xi}^k_{\varphi_y y}, \quad (38)$$

where only the left hand side depends on $\hat{\rho}_y^k$. In addition, (38) is affine in $\hat{\rho}_y^k$, hence it can be re-expressed as

$$H_k \hat{\rho}_y^k = h_k, \quad (39)$$

where $H_k$ is a $n_a \times n_a+1$ matrix built up from elements of $\hat{a}_k$ and $h_k$ is a vector of length $n_a$ given by the right hand side of (38). This is a system of equations with more unknowns than equations, but the set of all possible solutions can be formalised as

$$\rho_y^k = \alpha_k N(H_k) + H_k^{\dagger} h_k, \quad (40)$$

where $N(\cdot)$ denotes the nullspace and $\alpha_k$ is a scalar factor. It is necessary to distinguish between the input measurement noise variance obtained by (35) in step 1, and the quantity which would be obtained by the Frisch equation (20). Therefore, introduce

$$\hat{\varsigma}_k \triangleq \lambda_{\min}\left(B_k(\alpha_k)\right), \quad (41)$$

where the matrix $B_k$ is now a function of $\alpha_k$. Using (41) it is possible to search for that $\alpha_k$ which is in best agreement with the previously determined $\hat{\sigma}^k_{\tilde{u}}$, i.e.

$$\hat{\alpha}_k = \arg\min_{\alpha_k} ||J_k||_2^2, \quad (42)$$

where the cost function

$$J_k \triangleq \hat{\sigma}_{\tilde{u}}^k - \hat{\zeta}_k \tag{43}$$

measures the distance between the input noise variance estimate $\hat{\sigma}_{\tilde{u}}^k$ determined in Step 1 and the input noise variance estimate $\hat{\zeta}_k$ which is obtained using the $n_a$ normal equations (37) together with the Frisch equation (41) depending on the choice of $\alpha_k$. Once $\hat{\alpha}_k$ is determined, it is substituted in (40) to obtain $\hat{\rho}_y^k$, the searched estimate of the auto-covariance elements of the coloured output measurement noise $\tilde{y}_k$.

# 4 RECURSIVE SCHEME

## 4.1 Step 1

### 4.1.1 Recursive Update of Covariance Matrices

In order to satisfy the requirements of a recursive algorithm to store all data in a finite dimensional vector, the covariance matrices are updated via

$$\hat{\Sigma}_{\varphi}^k = \hat{\Sigma}_{\varphi}^{k-1} + \gamma_k \left( \overline{\varphi}_k \overline{\varphi}_k^T - \hat{\Sigma}_{\varphi}^{k-1} \right), \quad \hat{\Sigma}_{\varphi}^0 = 0, \tag{44a}$$

$$\hat{\Sigma}_{\overline{\zeta}\varphi}^k = \hat{\Sigma}_{\overline{\zeta}\varphi}^{k-1} + \gamma_k \left( \overline{\zeta}_k \overline{\varphi}_k^T - \hat{\Sigma}_{\overline{\zeta}\varphi}^{k-1} \right), \quad \hat{\Sigma}_{\overline{\zeta}\varphi}^0 = 0, \tag{44b}$$

where the normalising gain $\gamma_k$ is given by

$$\gamma_k \triangleq \frac{\gamma_{k-1}}{\lambda + \gamma_{k-1}}, \qquad \gamma_0 = 1 \tag{45}$$

with $0 < \lambda \leq 1$ being the forgetting factor giving exponential forgetting. From (44), the block matrices required in (28) and (37) are readily obtained.

### 4.1.2 Recursive Update of $\hat{\sigma}_{\tilde{u}}^k$

For the determination of $\hat{\sigma}_{\tilde{u}}^k$, an iterative optimisation procedure can be utilised to minimise (36) where it is iterated once at each step, leading to a recursive scheme (Ljung and Söderström, 1983; Ljung, 1999). Here, an iterative Newton method is utilised for this purpose, however other choices are also possible. The Newton method given by (Ljung, 1999, p. 326) is

$$\sigma_{\tilde{u}}^k = \sigma_{\tilde{u}}^{k-1} - \left[ V_k'' \right]^{-1} V_k', \tag{46}$$

where $V_k'$ and $V_k''$ denote the first and second order derivative of $V_k$ with respect to $\sigma_{\tilde{u}}^k$ evaluated at $\sigma_{\tilde{u}}^{k-1}$. The formulas for the derivatives are given in Appendices A and B, respectively.

**Remark 1.** *In order to stabilise the algorithm, it might be advantageous to restrict the search for the input measurement noise variance to the interval*

$$0 \leq \sigma_{\tilde{u}} \leq \sigma_{\tilde{u}}^{max}, \tag{47}$$

*where $\sigma_{\tilde{u}}^{max}$ is the maximal admissible value for $\sigma_{\tilde{u}}$, which can be computed from the data as discussed in (Beghelli et al., 1990). Alternatively, a positive constant can be chosen for the maximum admissible value, if such a-priori knowledge is available.*

### 4.1.3 Recursive Update of $\hat{\theta}_k$

In order to obtain a recursive expression for $\hat{\theta}_k$, an approach is adopted here, similar to that in (Ding et al., 2006), where the bias of the recursive LS estimate is compensated at each time step $k$.

Ignoring the influence of $\hat{\sigma}_{\tilde{u}}^k$ in (28), the uncompensated overdetermined IV normal equations can be expressed as

$$\frac{1}{k} \sum_{i=1}^{k} \begin{bmatrix} \varphi_{u_i} \\ \zeta_i \end{bmatrix} \begin{bmatrix} \varphi_{y_i}^T & \varphi_{u_i}^T \end{bmatrix} \hat{\theta}_k^{\mathrm{IV}} = \frac{1}{k} \sum_{i=1}^{k} \begin{bmatrix} \varphi_{u_i} \\ \zeta_i \end{bmatrix} y_i, \tag{48}$$

where $\hat{\theta}_k^{\mathrm{IV}}$ denotes the uncompensated (biased) estimate of $\theta$. Since one unknown, namely $\hat{\sigma}_{\tilde{u}}^k$, has already been obtained, it is sufficient to consider $n_a + n_b$ equations only[1], by disregarding the last equation of (48). Thus the uncompensated IV estimate is given as

$$\hat{\theta}_k^{\mathrm{IV}} = \left[ \frac{1}{k} \sum_{i=1}^{k} \delta_i \varphi_i^T \right]^{-1} \frac{1}{k} \sum_{i=1}^{k} \delta_i y_i, \tag{49}$$

where $\delta_i$ is obtained by deleting the last entry of $\overline{\delta}_i$. In order to obtain an explicit expression for the bias, the linear regression formulation

$$y_i = \varphi_i^T \theta + e_i \tag{50}$$

is substituted in (49) which gives

$$\begin{aligned} \hat{\theta}_k^{\mathrm{IV}} &= \left[ \frac{1}{k} \sum_{i=1}^{k} \delta_i \varphi_i^T \right]^{-1} \frac{1}{k} \sum_{i=1}^{k} \delta_i \left( \varphi_i^T \theta + e_i \right) \\ &= \theta + \left[ \frac{1}{k} \sum_{i=1}^{k} \delta_i \varphi_i^T \right]^{-1} \frac{1}{k} \sum_{i=1}^{k} \delta_i e_i \end{aligned} \tag{51}$$

By substituting $e_i = -\tilde{\varphi}_i \theta + \tilde{y}_i$ it follows that

$$\begin{aligned} \hat{\theta}_k^{\mathrm{IV}} = \theta &+ \left[ \frac{1}{k} \sum_{i=1}^{k} \delta_i \varphi_i^T \right]^{-1} \frac{1}{k} \sum_{i=1}^{k} \delta_i \tilde{y}_i \\ &- \left[ \frac{1}{k} \sum_{i=1}^{k} \delta_i \varphi_i^T \right]^{-1} \frac{1}{k} \sum_{i=1}^{k} \delta_i \tilde{\varphi}_i^T \theta. \end{aligned} \tag{52}$$

The vector $\delta_i$ is uncorrelated with $\tilde{y}_i$ which means that the middle part of the sum in (52) diminishes in the

[1] This corresponds to a basic IV estimator where the number of unknowns is equal to the length of the instrument vector.

asymptotic case, whereas

$$\lim_{k\to\infty} \quad \frac{1}{k}\sum_{i=1}^{k}\begin{bmatrix}\varphi_{u_i}\\ \zeta_i\end{bmatrix}\begin{bmatrix}\tilde{\varphi}_{y_i}^T & \tilde{\varphi}_{u_i}^T\end{bmatrix}^T = \begin{bmatrix}0 & \sigma_{\tilde{u}} I_{n_b}\\ 0 & 0\end{bmatrix}. \quad (53)$$

Consequently, for $k \to \infty$ (52) becomes

$$\theta^{\mathrm{IV}} = \theta - \sigma_{\tilde{u}}\Sigma_{\delta\varphi}^{-1} J\theta, \quad (54)$$

where $J$ is obtained by deleting the last row of $\bar{J}$ in (31). Equation (54) gives rise to the recursive bias compensation update equation for $\hat{\theta}_k$

$$\hat{\theta}_k = \hat{\theta}_k^{\mathrm{IV}} + \hat{\sigma}_{\tilde{u}}^k \left[\hat{\Sigma}_{\delta\varphi}^k\right]^{-1} J\hat{\theta}_{k-1}, \quad (55)$$

where the uncompensated parameter estimate $\hat{\theta}_k^{\mathrm{IV}}$ can be recursively computed via a recursive IV (RIV) algorithm (Ljung, 1999, p. 369) given by

$$\hat{\theta}_k^{\mathrm{IV}} = \hat{\theta}_{k-1}^{\mathrm{IV}} + L_k\left[y_k - \varphi_k^T\hat{\theta}_{k-1}^{\mathrm{IV}}\right], \quad (56a)$$

$$L_k = \frac{P_{k-1}\delta_k}{\frac{1-\gamma_k}{\gamma_k} + \varphi_k^T P_{k-1}\delta_k}, \quad (56b)$$

$$P_k = \frac{1}{1-\gamma_k}\left[P_{k-1} - \frac{P_{k-1}\delta_k\varphi_k^T P_{k-1}}{\frac{1-\gamma_k}{\gamma_k} + \varphi_k^T P_{k-1}\delta_k}\right]. \quad (56c)$$

with the only difference being that $P_k$ is scaled such that

$$\left[\hat{\Sigma}_{\delta\varphi}^k\right]^{-1} = P_k. \quad (57)$$

This avoids the matrix inversion in (55) by substituting (57) in (55).

### 4.2 Step 2

In order to solve (42) recursively, the Newton method is applied where it is iterated once as new data arrives. Consequently, the first and second order derivative of the cost function $J_k$ in (43) are to be determined w.r.t. $\alpha_k$, which are denoted $J_k'$ and $J_k''$, respectively. These are given by

$$J_k' = -2\left(\hat{\sigma}_{\tilde{u}}^k - \hat{\varsigma}_k\right)\hat{\varsigma}_k', \quad (58a)$$

$$J_k'' = \hat{\varsigma}_k', \quad (58b)$$

where $\hat{\varsigma}_k'$ denotes the derivative of $\hat{\varsigma}_k$ w.r.t. $\alpha_k$ and for which an approximation is derived in Appendix C. The recursive update for $\hat{\alpha}_k$ is therefore given by

$$\hat{\alpha}_k = \hat{\alpha}_{k-1} - \left[J_k''\right]^{-1} J_k', \quad (59)$$

whilst

$$\hat{\rho}_y^k = \hat{\alpha}_k N(H_k) + H_k^{\dagger} h_k. \quad (60)$$

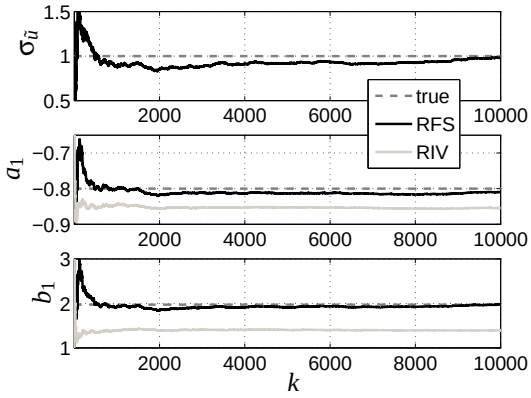



Figure 2: Recursive estimates for θ and $\sigma_{\tilde{u}}$ using the recursive Frisch scheme (RFS) and the biased recursive instrumental-variables (RIV) solution of the uncompensated normal equations.

## 5 SIMULATION

To compare the results of the recursive Frisch scheme (RFS) with the non-recursive algorithm, the system is chosen similar to that of Example 2 in (Söderström, 2008), i.e. a LTI SISO system with $n_a = n_b = 1$, and characterised, using (12), by

$$\vartheta = \begin{bmatrix}-0.8 & 2 & 1 & 1.96 & 1.37\end{bmatrix}^T. \quad (61)$$

The values for $r_{\tilde{y}}(0)$ and $r_{\tilde{y}}(1)$ arise by generating the output noise by the auto-regressive model

$$\tilde{y}_k = \frac{1}{1-0.7q^{-1}}v_k, \quad (62)$$

where $v_k$ is a zero-mean white process with unity variance. The system is simulated for $10,000$ samples using a zero mean, white and Gaussian distributed input signal of unity variance. The corresponding signal-to-noise ratio for input and output is given by 10.60dB and 39.12dB, respectively.

Choosing $\lambda = 1$, the results for Step 1 are shown in Figure 2. The first subplot shows that the Newton method is able to recursively estimate the input measurement noise variance $\sigma_{\tilde{u}}$. The remaining two subplots compare the RIV solution $\hat{\theta}_k^{\mathrm{IV}}$ of the uncompensated normal equations with the recursively compensated Frisch scheme estimate $\hat{\theta}_k$. As expected, the RIV is biased whilst the the RFS successfully compensates for this.

Figure 3 shows the estimates of $\rho_y$ obtained in Step 2 for both the RFS as well as the off-line case. In contrast to the results obtained in Step 1, the quality of the estimates obtained in Step 2 for $\hat{\rho}_y^k$ is inferior. This is in agreement with the results reported in (Söderström, 2008), where a Monte-Carlo

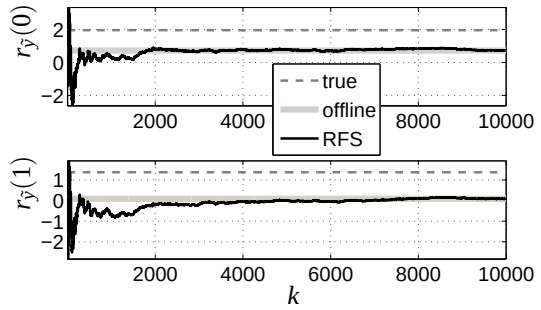



Figure 3: Recursive estimates for $r_{\tilde{y}}(0)$ and $r_{\tilde{y}}(1)$ using the recursive Frisch scheme (RFS).

analysis shows poor performance for $\hat{\rho}_y^k$ in the non-recursive case. The important observation to note here is that the recursively obtained estimates of $r_{\tilde{y}}(0)$ and $r_{\tilde{y}}(0)$ coincide with their off-line counterparts after $k = 10{,}000$ recursions. It is also observed that the values of $\hat{r}_{\tilde{y}}^k(0)$ (the estimated variance of the output measurement noise) occasionally exhibits a negative sign during the first 500 recursion steps. This could be avoided by projecting the estimates, such that

$$0 < \hat{\Sigma}_{\bar{\varphi}_y}^k < \hat{\Sigma}_{\bar{\varphi}_y}^k - \hat{\Sigma}_{\varphi_y\varphi_u}^k \left[\hat{\Sigma}_{\varphi_u}^k\right]^{-1} \hat{\Sigma}_{\varphi_u\bar{\varphi}_y}^k \tag{63}$$

is satisfied (Söderström, 2008).

# 6 CONCLUSIONS

The Frisch scheme for the coloured output noise case has been reviewed and a recursive algorithm for its adaptive implementation has been developed. The parameter vector is estimated utilising a recursive bias-compensating instrumental variables approach, where the bias is compensated at each time step. The input measurement noise variance and the output measurement noise auto-covariance elements are obtained via two (distinct) Newton algorithms. A simulation study illustrates the performance of the proposed algorithm.

Further work could concern computational aspects of the algorithm as well as its extension to the bilinear case.

# REFERENCES

Beghelli, S., Guidorzi, R. P., and Soverini, U. (1990). The Frisch scheme in dynamic system identification. *Automatica*, 26(1):171–176.

Ding, F., Chen, T., and Qiu, L. (2006). Bias compensation based recursive least-squares identification algorithm for MISO systems. *IEEE Trans. on Circuits and Systems*, 53(5):349–353.

Diversi, R., Guidorzi, R., and Soverini, U. (2003). Algorithms for optimal errors-in-variables filtering. *Systems & Control Letters*, 48:1–13.

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.

Linden, J. G., Vinsonneau, B., and Burnham, K. J. (2007). Fast algorithms for recursive Frisch scheme system identification. In *Proc. CD-ROM IAR & ACD Int. Conf.*, Grenoble, France.

Linden, J. G., Vinsonneau, B., and Burnham, K. J. (2008). Gradient-based approaches for recursive Frisch scheme identification. To be published at the 17th IFAC World Congress 2008.

Ljung, L. (1999). *System Identification - Theory for the user*. PTR Prentice Hall Infromation and System Sciences Series. Prentice Hall, New Jersey, 2nd edition.

Ljung, L. and Söderström, T. (1983). *Theory and Practice of Recursive Identification*. M.I.T. Press, Cambridge, MA.

Söderström, T. (2007a). Accuracy analysis of the Frisch scheme for identifying errors-in-variables systems. *Automatica*, 52(6):985–997.

Söderström, T. (2007b). Errors-in-variables methods in system identification. *Automatica*, 43(6):939–958.

Söderström, T. (2008). Extending the Frisch scheme for errors-in-variables identification to correlated output noise. *Int. J. of Adaptive Contr0l and Signal Proc.*, 22(1):55–73.

# APPENDIX

## A First Order Derivative of $V_k$

Denoting $(\cdot)'$ the derivative w.r.t. $\hat{\sigma}_{\tilde{u}}^k$ and introducing

$$f_k \triangleq \hat{G}_k^T \hat{\xi}_{\bar{\delta}y}^k, \qquad F_k \triangleq \hat{G}_k^T \hat{G}_k, \tag{64}$$

it holds that

$$f_k' = -\begin{bmatrix} 0 \\ \hat{\xi}_{\varphi_u y}^k \end{bmatrix}, \tag{65a}$$

$$F_k' = \begin{bmatrix} 0 & \hat{\Sigma}_{\varphi_u\varphi_y}^{k\,T} \\ \hat{\Sigma}_{\varphi_u\varphi_y}^k & 2\hat{\sigma}_{\tilde{u}}^{k-1} I_{n_b} - 2\hat{\Sigma}_{\varphi_u}^k \end{bmatrix}, \tag{65b}$$

$$F_k^{-1\,\prime} = -F_k^{-1} F_k' F_k^{-1} \tag{65c}$$

and the first order derivative is given by

$$\begin{aligned} V_k' &= -\left(f_k^T F_k^{-1} f_k\right)' \\ &= -f_k'^{\,T} F_k^{-1} f_k - f_k^T F_k^{-1\,\prime} f_k - f_k^T F_k^{-1} f_k'. \end{aligned} \tag{66}$$

## B Second Order Derivative of $V_k$

Utilising the product rule, the second order derivative is given by

$$\begin{aligned} V_k'' &= -\,{f_k'}^T {F_k^{-1}}' f_k - {f_k'}^T F_k^{-1} f_k' \\ &\quad -\,{f_k'}^T {F_k^{-1}}' f_k - f_k^T {F_k^{-1}}'' f_k - f_k^T {F_k^{-1}}' f_k' \\ &\quad -\,{f_k'}^T F_k^{-1} f_k' - f_k^T {F_k^{-1}}' f_k' \\ &= -2\,{f_k'}^T {F_k^{-1}}' f_k - 2\,{f_k'}^T F_k^{-1} f_k' \\ &\quad - f_k^T {F_k^{-1}}'' f_k - 2 f_k^T {F_k^{-1}}' f_k' \end{aligned} \tag{67}$$

with

$${F_k^{-1}}'' = -{F_k^{-1}}' F_k' F_k^{-1} - F_k^{-1} F_k'' F_k^{-1} - F_k^{-1} F_k' {F_k^{-1}}', \tag{68a}$$

$$F_k'' = \begin{bmatrix} 0 & 0 \\ 0 & 2I_{n_b} \end{bmatrix}. \tag{68b}$$

## C Derivative of $\hat{\varsigma}_k$

The idea is to linearise the Frisch equation (20) using perturbation theory, in order to approximate the derivative of $\hat{\varsigma}_k$ w.r.t. $\alpha_k$. The derivation here is conceptually similar to that given in Appendix II.B of (Söderström, 2007a), but with the linearisation carried out around $\hat{\vartheta}_{k-1}$ rather than the 'true' parameters.

Assume that at time instance $k-1$, $\hat{\vartheta}_{k-1}$ satisfies the extended compensated normal equations

$$\begin{bmatrix} \hat{\Sigma}^{k-1}_{\bar{\varphi}_y} - \hat{\Sigma}^{k-1}_{\bar{\varphi}_y} & \hat{\Sigma}^{k-1}_{\bar{\varphi}_y \varphi_u} \\ \hat{\Sigma}^{k-1}_{\varphi_u \bar{\varphi}_y} & \hat{\Sigma}^{k-1}_{\varphi_u} - \hat{\sigma}^{k-1}_{\tilde{u}} I_{n_b} \end{bmatrix} \begin{bmatrix} \hat{\tilde{a}}_{k-1} \\ \hat{b}_{k-1} \end{bmatrix} = 0 \tag{69}$$

which are rewritten for ease of notation as

$$\begin{bmatrix} \mathfrak{A} - \mathfrak{B} & \mathfrak{C} \\ \mathfrak{C}^T & \mathfrak{D} - \hat{\sigma}^{k-1}_{\tilde{u}} I \end{bmatrix} \begin{bmatrix} \mathfrak{a} \\ \mathfrak{b} \end{bmatrix} = 0. \tag{70}$$

Similarly, introduce the notation at time instance $k$ as

$$\begin{bmatrix} \mathcal{A} - \mathcal{B} & \mathcal{C} \\ \mathcal{C}^T & \mathcal{D} - \hat{\sigma}^{k}_{\tilde{u}} I \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = 0. \tag{71}$$

Let $\hat{\sigma}^k_{\tilde{u}}$ denote the estimate obtained via (35). Alternatively, if $\hat{\Sigma}^k_{\bar{\varphi}_y}$ is known, the input measurement noise could be obtained using (20) and denote this quantity $\varsigma^k$. Using perturbation theory for eigenvalues yields

$$\begin{aligned} \varsigma^k &= \lambda_{\min}\{B_k(\alpha_k)\} = \lambda_{\min}\{B_{k-1}(\alpha_{k-1}) + \Delta B_k\} \\ &\approx \varsigma^{k-1} + \frac{\mathfrak{b}^T \Delta B_k \mathfrak{b}}{\mathfrak{b}^T \mathfrak{b}}, \end{aligned} \tag{72}$$

where the perturbation is given by (cf. (21))

$$\begin{aligned} \Delta B_k &= B_k(\alpha_k) - B_{k-1}(\alpha_{k-1}) \\ &= \mathcal{D} - \mathcal{C}^T [\mathcal{A} - \mathcal{B}]^{-1} \mathcal{C} - \mathfrak{D} + \mathfrak{C}^T [\mathfrak{A} - \mathfrak{B}]^{-1} \mathfrak{C} \\ &= \mathcal{D} - \mathcal{C}^T {\mathcal{F}}^{-1} \mathcal{C} - \mathfrak{D} + \mathfrak{C}^T \mathfrak{F}^{-1} \mathfrak{C} \end{aligned} \tag{73}$$

with $\mathcal{F} \triangleq [\mathcal{A} - \mathcal{B}]$ and $\mathfrak{F} \triangleq [\mathfrak{A} - \mathfrak{B}]$. Substituting (73) in (72) yields

$$\begin{aligned} \varsigma^k - \varsigma^{k-1} &\approx \frac{\mathfrak{b}^T}{\mathfrak{b}^T \mathfrak{b}} \left( \mathcal{D} - \mathfrak{D} + \mathfrak{C}^T \mathfrak{F}^{-1} \mathfrak{C} - \mathcal{C}^T {\mathcal{F}}^{-1} \mathcal{C} \right) \mathfrak{b} \\ &= \frac{\mathfrak{b}^T}{\mathfrak{b}^T \mathfrak{b}} (\mathcal{D} - \mathfrak{D}) \mathfrak{b} + \frac{\mathfrak{b}^T X \mathfrak{b}}{\mathfrak{b}^T \mathfrak{b}}, \end{aligned} \tag{74}$$

where $X$ can be expressed as

$$\begin{aligned} X &= \mathfrak{C}^T \mathfrak{F}^{-1} \mathfrak{C} - \mathcal{C}^T {\mathcal{F}}^{-1} \mathcal{C} \\ &\quad + \mathfrak{C}^T {\mathcal{F}}^{-1} \mathcal{C} - \mathfrak{C}^T {\mathcal{F}}^{-1} \mathcal{C} \\ &\quad + \mathfrak{C}^T \mathfrak{F}^{-1} \mathcal{C} - \mathfrak{C}^T \mathfrak{F}^{-1} \mathcal{C} \\ &= \left( \mathfrak{C}^T - \mathcal{C}^T \right) {\mathcal{F}}^{-1} \mathcal{C} + \mathfrak{C}^T \mathfrak{F}^{-1} (\mathfrak{C} - \mathcal{C}) \\ &\quad - \mathfrak{C}^T \mathfrak{F}^{-1} (\mathfrak{F} - \mathcal{F}) {\mathcal{F}}^{-1} \mathcal{C} \end{aligned} \tag{75}$$

and by combining (74) and (75), it holds that

$$\begin{aligned} \mathfrak{b}^T \mathfrak{b} \left( \varsigma^k - \varsigma^{k-1} \right) &\approx \mathfrak{b}^T (\mathcal{D} - \mathfrak{D}) \mathfrak{b} \\ &\quad + \mathfrak{b}^T \left( \mathfrak{C}^T - \mathcal{C}^T \right) {\mathcal{F}}^{-1} \mathcal{C} \mathfrak{b} \\ &\quad + \mathfrak{b}^T \mathfrak{C}^T \mathfrak{F}^{-1} (\mathfrak{C} - \mathcal{C}) \mathfrak{b} \\ &\quad - \mathfrak{b}^T \mathfrak{C}^T \mathfrak{F}^{-1} (\mathfrak{F} - \mathcal{F}) {\mathcal{F}}^{-1} \mathcal{C} \mathfrak{b}. \end{aligned} \tag{76}$$

Now, the first row of (70) gives

$$\mathfrak{a} = -\mathfrak{F}^{-1} \mathfrak{C} \mathfrak{b} \tag{77}$$

and by assuming that ${\mathcal{F}}^{-1} \mathcal{C} \mathfrak{b} \approx -\mathfrak{a}$, (76) finally simplifies to

$$\begin{aligned} \mathfrak{b}^T \mathfrak{b} \left( \varsigma^k - \varsigma^{k-1} \right) &\approx \mathfrak{b}^T (\mathcal{D} - \mathfrak{D}) \mathfrak{b} \\ &\quad - \mathfrak{b}^T \left( \mathfrak{C}^T - \mathcal{C}^T \right) \mathfrak{a} \\ &\quad - \mathfrak{a}^T (\mathfrak{C} - \mathcal{C}) \mathfrak{b} \\ &\quad - \mathfrak{a}^T (\mathfrak{F} - \mathcal{F}) \mathfrak{a}, \end{aligned} \tag{78}$$

where $\mathcal{F}$ is the only element depending on $\alpha_k$. Therefore,

$$\frac{d\varsigma^k}{d\alpha_k} \approx \frac{d}{d\alpha_k} \left( \frac{\mathfrak{a}^T (\mathcal{A} - \mathcal{B}) \mathfrak{a}}{\mathfrak{b}^T \mathfrak{b}} \right) = -\frac{\mathfrak{a}^T \frac{d\mathcal{B}}{d\alpha_k} \mathfrak{a}}{\mathfrak{b}^T \mathfrak{b}} \tag{79}$$

or equivalently

$$\frac{d}{d\alpha_k} \lambda_{\min}\{B_k(\alpha_k)\} \approx -\frac{\tilde{\hat{a}}^T_{k-1}}{\hat{b}^T_{k-1} \hat{b}_{k-1}} \frac{d}{d\alpha_k} \hat{\Sigma}^k_{\bar{\varphi}_y} \tilde{\hat{a}}_{k-1}. \tag{80}$$

Since $\Sigma^k_{\bar{\varphi}_y}$ consists of the quantities $\hat{r}^k_{\hat{y}}(0), \ldots, \hat{r}^k_{\hat{y}}(n_a)$, it remains to determine

$$\frac{d}{d\alpha_k} \hat{\rho}^k_y = \begin{bmatrix} \frac{d}{d\alpha_k} \hat{r}^k_{\hat{y}}(0) & \cdots & \frac{d}{d\alpha_k} \hat{r}^k_{\hat{y}}(n_a) \end{bmatrix}^T \tag{81}$$

which, due to (40), is given by

$$\frac{d}{d\alpha_k} \hat{\rho}^k_y = N(H_k). \tag{82}$$

# DISCRETE-EVENT SIMULATION OF A COMPLEX INTERMODAL CONTAINER TERMINAL

## *A Case-Study of Standard Unloading/Loading Processes of Vessel Ships*

Guido Maione
*DEESD, Technical University of Bari, Viale del Turismo 8, I-74100, Taranto, Italy*
*gmaione@poliba.it*

Keywords: Container Terminals, Discrete-Event Systems, Simulation, Transport Systems.

Abstract: This paper analyzes the performance of a complex maritime intermodal container terminal. The aim is to propose changes in the system resources or in handling procedures that guarantee better performance in perturbed conditions. A discrete-event system simulation study shows that, in future conditions of increased traffic volumes and reduced available stacking space, more internal transport vehicles, or appropriate scheduling and routing policies, or an increased degree of automation would improve the performance.

# 1 INTRODUCTION

In an intermodal container terminal (CT) freight is organized, stacked, handled and transported in standard units of a typical container, which is called TEU (Twenty Equivalent Unit) and which fits to ships, trains and trucks that are built and work for it.

A maritime CT is usually managed to offer three main services: a railway/road 'export cycle', when TEUs arrive by trains/trucks and depart on vessel ships; a railway/road 'import cycle', when TEUs arrive on vessel ships and depart by trains/trucks; a 'transshipment cycle', when TEUs arrive on vessel (feeder) ships and depart on feeder (vessel) ships. The hub in Taranto is managed by a private company (Taranto Container Terminal or TCT), whose primary business is for transhipment, because of the low quality of railway and road networks connecting the hub to Italy and the rest of Europe.

The terminal receives ships to a quay and uses yard blocks to stack full or empty TEUs. Imported TEUs are unloaded, exported TEUs are loaded, while in transshipped TEUs both processes occur. Full TEUs may be imported, exported or transshipped. Empty TEUs are unloaded from feeder ships or arrive on trains or trucks; then they are loaded on vessel ships. So, they are transhipped or exported.

The typical activities executed by humans and resources in a transshipment cycle are the following:

- Unloading TEUs from ship by quay cranes;
- Picking-up and transferring TEUs to a yard block by trailers;
- Picking-up and stacking TEUs in a yard block by yard cranes;
- Redistributing TEUs in yard blocks by yard cranes and trailers;
- Picking-up and transferring TEUs to ship by yard cranes and trailers;
- Loading on ship by quay cranes.

Managing these activities requires an optimized use of equipment and human operators. Human supervision is often required to control processes concurring and competing for the limited number of available resources. Moreover, efficiency is needed for services in reduced time without excessive costs: both the TCT needs to profitably use resources, and ship companies aim at saving the berthing time/cost.

TCT is expecting a growth in freight volumes and has recently expanded the yard. But no investment was made on local land infrastructures. Not much research was carried out on use of information and communication technologies or new control policies to improve efficiency, to the best of the author knowledge. Improvements can be achieved for TCT, which is very sensitive to disturbances and parameter variations (sudden or big increase of traffic volumes, reduction or reorganization of yard, changes in berthing spaces, different routing of trailers, faults and malfunctions).

Then, an intelligent control may guarantee robustness and a quick reaction to parameter variations. The aim here is to prove that current

organization and control of the main unloading and loading processes could be not efficient in future operating conditions. Changing management of operations is necessary to guarantee good performance in perturbed conditions.

# 2 LITERATURE OVERVIEW

Managing a maritime CT is a complex task. Several analytical models have been proposed as tool for the simulation of terminals useful to an optimal design and layout, organization, management and control.

Modelling CTs requires the simulation of many operations that need coordination to minimize time and costs. Determining the best management and control policies is also important (Mastrolilli *et al.*, 1998). The main problems are: berth allocation; loading and unloading of ships (crane assignment, stowage planning); transfer of TEUs from ships to yard and back; stacking operations; transfer to/from other transport modes; workforce scheduling.

A thorough literature review on modelling approaches is given in (Steenken *et al.*, 2004). Two main classes of modelling approaches can be highlighted: microscopic and macroscopic methods (Cantarella *et al.*, 2006). Microscopic models are generally based on discrete-event system simulation that may include Petri Nets (Fischer and Kemper, 2000, Liu and Ioannou, 2002), object-oriented (Bielli *et al.*, 2006) and queuing networks theory approaches (Legato and Mazza, 2001). Even if high computational effort may be required, microscopic simulation explicitly models all activities as well as the whole system by considering the single TEUs as entities. Then, it estimates performance as consequence of different designs and/or management scenarios.

Macroscopic modelling (de Luca *et al.*, 2005) is suitable for supporting strategic decisions, system design and layout, investments on handling equipment. A network-based approach is presented in (Kozan, 2000) for optimising efficiency by using a linear programming method.

# 3 DEVS MODELLING

A Discrete EVent System (DEVS) specification technique (Zeigler *et al.,* 2000) completely and unambiguously represents and controls the terminal processes.

Atomic dynamic DEVSs model both TEUs flowing in the system and resources (cranes, trailers, trucks) used to handle them. DEVSs interact by transmitting outputs and receiving inputs, which are all instantaneous events. Timed processes are defined by a start-event and a stop-event.

For each DEVS, internal events are triggered by internal mechanisms, external input events are determined by other DEVSs, and external output events are generated and directed to other entities.

A DEVS state is changed by an input or when the time specified before an internal event elapses. In the first case, an external transition function determines the state next to the received input; in the second case, an internal transition function gives the state next to the internal event. The total state is $\mathbf{q} = (\mathbf{s}, e)$, where $\mathbf{s}$ is the sequential state and $e$ is the time elapsed since the last transition.

To summarize, each DEVS is represented as:

$$\boldsymbol{DEVS} = < \mathbf{X}, \mathbf{Y}, \mathbf{S}, \delta_{int}, \delta_{ext}, \lambda, ta > \quad (1)$$

where $\mathbf{X}$ is the set of inputs, $\mathbf{Y}$ is the set of outputs, $\mathbf{S}$ is the set of sequential states, $\delta_{int}:\mathbf{S}\rightarrow\mathbf{S}$ is the internal transition function, $\delta_{ext}:\mathbf{Q}\times\mathbf{X}\rightarrow\mathbf{S}$ is the external transition function, $\mathbf{Q}=\{\mathbf{q}=(\mathbf{s},e)|\mathbf{s}\in\mathbf{S},0\leq e\leq ta(\mathbf{s})\}$, $\lambda:\mathbf{S}\rightarrow\mathbf{Y}$ is the output function, $ta:\mathbf{S}\rightarrow\Re_0^+$ is the time advance function, with $\Re_0^+$ set of positive real numbers with 0 included.

The network of DEVS atomic models is used as a platform for simulating the TCT dynamics. Details are omitted here for sake of space.

# 4 SIMULATION ANALYSIS

A simulation study is presented to analyse the contemporaneous processes of unloading and loading TEUs from and to a vessel ship.

The simulation model is based on the real TCT equipment and operation times, which were statistically observed during steady-state conditions.

The model was developed in a discrete-event environment by using Arena® (Kelton *et al.*, 1998).

## 4.1 Experimental Setup

The data used to set up the simulation experiments refer to the observations recorded during year 2004, when TCT achieved the maximum productivity (Table 1). About 14% of TEUs flew through railway/road transport modes. The numbers of full/empty TEUs are divided as in Table 2.

Table 1: Loaded/unloaded TEUs in TCT (2004).

| Loaded TEUs | | Unloaded TEUs | |
|---|---|---|---|
| Full | 273224 | Full | 285488 |
| Empty | 108172 | Empty | 96434 |
| Total TL | 381396 | Total TD | 381922 |
| Total T = TL+TD = 763318 | | | |
| TEUs on railway 44486 ≅ 5.8% of T | | | |
| TEUs on road 64648 ≅ 8.5% of T | | | |

Table 2: Flows of containers in TCT (2004).

| Containers (Cycle) | No. |
|---|---|
| Full, from vessel to feeder | $x$ |
| Full, from feeder to vessel | $y$ |
| Full, from vessel to train/truck | $t$ |
| Full, from train/truck to vessel | $z$ |
| Empty, from feeder to blocks | $r$ |
| Empty, from train/truck to blocks | $h$ |
| Empty, from blocks to vessel | $q$ |

Then, we may establish the following relations:

$$x + y + z = 273224 \quad (2)$$
$$q = 108172 \quad (3)$$
$$x + y + t = 285488 \quad (4)$$
$$r = 96434 \quad (5)$$
$$t + z + h = 113134 \quad (6)$$
$$r + h = q \quad (7)$$

where (7) is due to the assumption that no empty TEU is accumulated and left in the yard blocks.

Then, it is easy to find: $x+y = 228658$, $t = 56830$, $z = 44566$, $r = 96434$, $h = 11738$, $q = 108172$. The TEUs separately handled by vessel and feeder ships were estimated in the ranges in Table 3, because $x$ and $y$ were assumed between 0 and 228658. Then, the average number of TEUs handled by vessel (*avs*) or feeder ships (*afs*) was determined by assuming traffic volumes of 346 vessel and 570 feeder ships in year 2004. These assumptions were based on the traffic data available for year 2003 and on the 15.9% increase in traffic (then in number of ships) in 2004.

Table 3: Containers handled by ships.

| Vessels | TEUs | Est. Range | *avs* |
|---|---|---|---|
| Unload. | $x+t$ | [56830,285488] | [164,825] |
| Loaded | $y+z+q$ | [152738,381396] | [441,1102] |
| Total | $x+t+y+z+q$ | 438226 | 1266 |
| Feeders | TEUs | Est. Range | *afs* |
| Unload. | $y+r$ | [96434,325092] | [169,570] |
| Loaded | $x$ | [0,228658] | [0,401] |
| Total | $y+r+x$ | 325092 | 570 |

If $x = y = 114329$, then the flows indicated by Tables 4 and 5 are obtained, which were used to set-up the simulation tests. Flows of TEUs from vessel ships to land are in a ratio 8 to 6 between road and railway modes, as observed in 2004. Unloaded and loaded TEUs are 39% and 61% of the total for vessel ships, 65% and 35% for feeder ships.

Table 4: Containers unloaded (U) and loaded (L) by vessel ships (F = feeder ships, TA = trains, TU = trucks, E = blocks for empty TEUs).

| U | No. (%) | L | No. (%) |
|---|---|---|---|
| To F | 114329 (66.80) | From F | 114329 (42.81) |
| To TA | 24356 (14.23) | From E | 108172 (40.50) |
| To TU | 32474 (18.97) | From TA | 19100 (7.15) |
| Total | 171159 (100) | From TU | 25466 (9.54) |
| | | Total | 267067 (100) |

Table 5: Containers unloaded (U) and loaded (L) by feeder ships (V = vessel ships, E = blocks for empty TEUs).

| U | No. (%) | L | No. |
|---|---|---|---|
| To V | 114329 (54.24) | From V | 114329 |
| To E | 96434 (45.76) | Total | 114329 |
| Total | 210763 (100) | | |

## 4.2 Simulation Assumptions

Simulation is based on the following assumptions:

- Only 1 vessel ship is berthed, full TEUs are unloaded, full and empty TEUs are loaded; 1300 TEUs are handled; 508 (39%) are unloaded, 792 (61%) are loaded, according to the percentage partitions shown in Table 4;
- The average values of handled TEUs in Table 3 is used, because information about daily movement or ship size was not available;
- Simulation is limited by the time necessary to end the unloading and loading processes;
- Transfers from/to the railway connection or the truck gate, are not considered;
- Operations length and distances travelled are measured in minutes and meters, respectively.

The model considers four quay cranes: QC1 and QC2 are for unloading, QC3 and QC4 for loading. Then, unloaded TEUs are stowed in ship sections different from those reserved for loaded TEUs, so that the processes are parallel. Sometimes cranes sequentially unload and load TEUs, depending on the stowage plan and on the destinations of TEUs.

A quay crane unloads/loads two TEUs on/from a trailer in eight steps (Table 6): S1) picking the first TEU from ship/trailer; S2) moving the crane with first TEU towards the trailer/ship; S3) releasing the first TEU on the trailer/ship; S4) moving the crane back to the ship/trailer; S5) picking the second TEU from ship/trailer; S6) moving the crane with second

TEU to the trailer/ship; S7) releasing the second TEU; S8) moving the crane back to the ship/trailer.

Table 6: Operation cycle of quay cranes.

| Step | Duration |
|---|---|
| S1 | Tria(0.4375,0.5,0.75) |
| S2 | 0.333 |
| S3 | Tria(0.4375,0.5,0.75) |
| S4 | 0.667 |
| S5 | Tria(0.4375,0.5,0.75) |
| S6 | 0.333 |
| S7 | Tria(0.4375,0.5,0.75) |
| S8 | 0.667 |

The triangular distribution is used because only the estimates of the minimum, most likely and maximum values (shown in this order) of the processing times are known. Simple translational return steps last longer (twice) than transfer steps because the crane is more unstable without TEUs.

Five trailers serve each quay crane. Each set of five trailers is indicated with a unique symbol: TR1, TR2, TR3, TR4 are associated to QC1, QC2, QC3, QC4, respectively. Each trailer always transports two TEUs, with a speed of 300 m/minutes (400 m/minutes when travelling unloaded). The closest trailer is selected for a task between ship and yard.

Before being loaded, exported and transshipped TEUs are stacked in blocks close to the quay area, while imported TEUs are stacked in blocks close to the land connections. Only one yard crane works on each block for unloaded TEUs from ships: YC1 serves transhipped TEUs; YC2/YC3 serves exported TEUs. Two yard cranes (YC4 and YC5 or YC6 and YC7) work for each block for TEUs to be loaded: YC4 serves empty TEUs, YC5 serves full TEUs; YC6 and YC7 serve full TEUs. YC7 has priority with respect to YC6 because it is closer to the quay. TEUs picked by YC4 and YC5 are loaded by QC3, those picked by YC6 and YC7 are loaded by QC4.

A yard crane unloads/loads two TEUs in/from a yard block from/to a trailer in eight steps (Table 7).

Table 7: Operation cycle of yard cranes.

| Step | Duration |
|---|---|
| S1 | Tria(0.125,0.375,0.625) |
| S2 | 0.25 |
| S3 | Tria(0.125,0.375,0.625) |
| S4 | 0.5 |
| S5 | Tria(0.125,0.375,0.625) |
| S6 | 0.25 |
| S7 | Tria(0.125,0.375,0.625) |
| S8 | 0.5 |

A typical and important performance index is the *Ship Turn-around Time* (STT), the average time spent by a berthed ship to unload and load TEUs. STT is measured between ship arrival and departure. Minimizing STT is the main objective of every terminal management. An empiric relation to calculate the minimum STT value is:

$$STT_{min} = nc / (ct \times nqc) \quad (8)$$

where $nc$ is the number of unloaded/loaded TEUs, $ct$ is the *cycle time* (the number of moves/hour of a quay crane), and $nqc$ is the number of quay cranes.

Equation (8) gives a reference for the terminal productivity. Namely, it does not consider the dependence of $ct$ on $nqc$, due to the interaction between $nqc$ and the handling capacity in the limited quay space, and the effects of internal transfers. Figure 1 gives the STT when $nc$ = 3400 and $ct$ = 42 hours$^{-1}$. Equation (8) can also be used to estimate the necessary $nqc$ to achieve a desired STT.

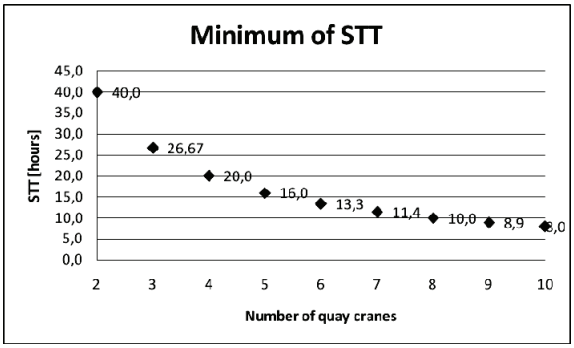



Figure 1: STT as function of $nqc$.

Assuming the most likely value $ct$ = 30 hours$^{-1}$, if $nc_u$ = 508 and $nc_l$ = 792 are the unloaded and loaded TEUs, and if $nqc_u = nqc_l = 2$ are the cranes used for the two processes, then a reference limit for STT is:

$$STT_{min} = \max\{ nc_u/(ct \times nqc_u); nc_l/(ct \times nqc_l)\} \quad (9)$$
$$= \max\{8.5; 13.2\} = 13.2 \text{ hours.}$$

Finally, note that performance is affected by the partial automation of processes, the humans' cooperation, the non-optimal ship distribution of TEUs and weather conditions.

## 4.3 Simulation Results

Ten simulation runs were executed, using different seeds for generating random variables, in order to obtain sufficient results for a statistical evaluation. Each run was terminated after 1300 unloaded and

loaded TEUs. The system state was initialized at the beginning of each run, to start from the same condition. Statistics were also initialized to have results independent on the data obtained from precedent runs. Initializations guarantee statistically independent and identically distributed replications of the terminating simulation.

STT was measured at the end of each run (Figure 2). The minimum, maximum, and average values were 891, 902, and 898, i.e. about 15 hours.

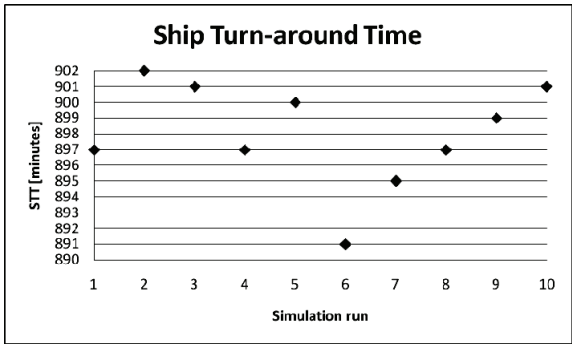



Figure 2: Measured STT in 10 simulation runs.

These results validate the model because:

- They are below the real TCT performance, because only 1 ship/day is served in standard real operating conditions;
- The measured values of STT are greater than the lower theoretical limit established by (9).

STT can be also measured for ships of different capacity or with a distribution of TEUs different from that in Table 4. If we let 1300 TEUs equally distributed between the four quay cranes, we obtain the results in Figure 3. The minimum, maximum and average values of STT were, respectively, 714, 723, and 718, that correspond to about 12 hours.

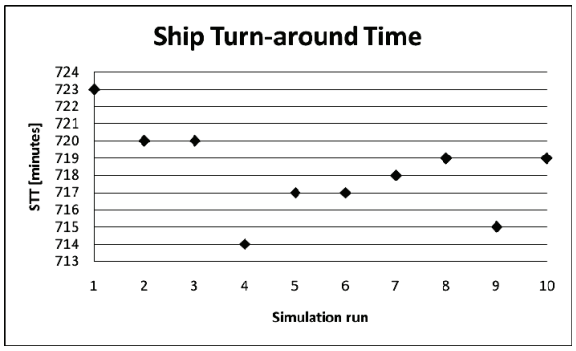



Figure 3: Measured STT in 10 simulation runs (TEUs equally distributed between quay cranes).

Performance indices were measured for critical resources like trailers and cranes: waiting times in queue; number of entities in queue; resource utilization. The associated statistics were: the average value in 10 runs; the minimum average value in a single run; the maximum average value in a single run; the maximum value.

Table 8 shows the waiting times. For unloading processes, TEUs may wait for the following busy resources: a) TR1 or TR2, when being on QC1 or QC2; b) YC1, YC2, YC3, when being on TR1 or TR2. For loading processes, TEUs may wait for: a) TR3, when being on YC4 used for empty TEUs; b) TR3, when being on YC5 used for full TEUs (busy resource TR3*); c) TR4, when being on YC6 used for full TEUs; d) TR4, when being on YC7 used for full TEUs (busy resource TR4*); e) QC3 or QC4, when being on trailers TR3 or TR4.

Table 8: Waiting times in queue of busy resources.

| Busy Res. | Average | Min. Aver. | Max. Aver. | Max. Value |
|---|---|---|---|---|
| TR1 | 0.0788 | 0.00 | 0.1393 | 5.0002 |
| TR2 | 0.0807 | 0.00 | 0.1613 | 4.7440 |
| YC1 | 3.5193 | 2.5625 | 4.6946 | 17.7338 |
| YC2 | 0.0612 | 0.00 | 0.1264 | 1.9246 |
| YC3 | 0.0989 | 0.00 | 0.1965 | 2.0696 |
| TR3 | 1.0662 | 1.0013 | 1.1135 | 6.8421 |
| TR3* | 28.4835 | 28.2937 | 28.6313 | 770.4400 |
| TR4 | 5.1725 | 5.1165 | 5.2450 | 418.2100 |
| TR4* | 1.2628 | 1.21478 | 1.3079 | 7.8171 |
| QC3 | 7.2326 | 7.0881 | 7.3267 | 10.5871 |
| QC4 | 5.9521 | 5.8538 | 6.0713 | 10.2105 |

The average waiting times of TR1 and TR2 are below 5 seconds, and then delays in unloading TEUs due to the waiting of trailers below the quay cranes can be neglected. So, more trailers are not necessary for unloading in the simulated conditions. On the contrary, the results for TR3, TR3*, TR4, TR4* show that the loading process waits for long time when yard cranes are used. Thus, at least one more trailer should be used.

The large values for TR3* and TR4 were obtained because of the priority given to empty with respect to full TEUs, and because of the priority of selecting the closest yard crane YC7 instead of YC6.

If we consider the interactions of trailers with yard cranes during the unloading process, high waiting times are observed for YC1 only, because most of the unloaded TEUs were stacked in the block served by YC1. More yard cranes would speed-up the stacking process, but they are not necessary since the number and speed of trailers is

sufficient to guarantee fast and almost continuous unloading operations by the quay cranes.

Long times are recorded for trailers when waiting for quay cranes to load TEUs (more than 7 minutes for QC3 and about 6 minutes for QC4). Then, one more trailer could help operations in the yard area, because the maximum number of queued trailers below a quay crane is three (see Table 9), such that the other two are available for yard cranes.

Table 9 shows the results for the number of entities in queue (the minimum value is always 0).

Table 10 shows the utilization of resources, i.e. the percentage number of busy units or the percentage busy time for single-unit resources (the minimum is always 0, the maximum is always 1).

Table 9: Number of entities in queue of busy resources.

| Busy Res. | Average | Min. Aver. | Max. Aver. | Max. Value |
|---|---|---|---|---|
| TR1 | 0.0111 | 0.00 | 0.0199 | 1.0000 |
| TR2 | 0.0114 | 0.00 | 0.0230 | 1.0000 |
| YC1 | 0.6708 | 0.4637 | 0.9167 | 6.0000 |
| YC2 | 0.0026 | 0.00 | 0.0062 | 1.0000 |
| YC3 | 0.0056 | 0.00 | 0.0109 | 1.0000 |
| TR3 | 0.2018 | 0.1911 | 0.2100 | 1.0000 |
| TR3* | 0.8882 | 0.8837 | 0.8906 | 1.0000 |
| TR4 | 0.5703 | 0.5653 | 0.5770 | 1.0000 |
| TR4* | 0.1392 | 0.1339 | 0.1446 | 1.0000 |
| QC3 | 1.5947 | 1.5753 | 1.6095 | 3.0000 |
| QC4 | 1.3124 | 1.2990 | 1.3337 | 3.0000 |

Table 10: Utilization of resources.

| Resource | Average | Min. Aver. | Max. Aver. |
|---|---|---|---|
| QC1 | 0.6104 | 0.6019 | 0.6218 |
| QC2 | 0.6122 | 0.6036 | 0.6265 |
| QC3 | 0.9919 | 0.9913 | 0.9926 |
| QC4 | 0.9378 | 0.9324 | 0.9439 |
| TR1 | 0.4818 | 0.4614 | 0.5013 |
| TR2 | 0.4828 | 0.4660 | 0.5049 |
| TR3 | 0.9897 | 0.9890 | 0.9899 |
| TR4 | 0.9360 | 0.9307 | 0.9425 |
| YC1 | 0.5712 | 0.5302 | 0.5970 |
| YC2 | 0.1150 | 0.0892 | 0.1692 |
| YC3 | 0.1617 | 0.1055 | 0.1812 |
| YC4 | 0.8642 | 0.8612 | 0.8724 |
| YC5 | 0.9825 | 0.9818 | 0.9831 |

Results for quay cranes indicate that unloading with QC1 and QC2 terminates before loading with QC3 and QC4. QC3 is used more than QC4 because of the high number of empty TEUs. Considerations about yard cranes are similar. Trailers TR1 and TR2 complete their tasks much earlier than TR3 and TR4, which are practically always busy. Then, the transport processes could benefit from more trailers.

# 5 CONCLUSIONS

This paper presents simulates a maritime terminal container (TCT) in standard operating conditions. Results prove the benefit from new control strategies different from those currently used. A new control approach could reduce terminal operating cycles in standard and, above all, in perturbed operating conditions.

# REFERENCES


Bielli, M., Boulmakoul, A., Rida, M., 2006. Object oriented model for container terminal distributed simulation. *European Journal of Operational Research*, Vol. 175, No. 3, pp. 1731-1751.

Cantarella, G.E., Cartenì, A., de Luca, S., 2006. A comparison of macroscopic and microscopic approaches for simulating container terminal operations. In *Proc. of EWGT2006 Joint conference*, Bari, Italy, 27-29 Sept. 2006.

de Luca, S., Cantarella, G.E., Cartenì, A., 2005. A macroscopic model of a container terminal based on diachronic networks. In *Proc. Second Workshop on the Schedule-Based Approach in Dynamic Transit Modelling*, Ischia, Naples, Italy, 29-30 May 2005.

Fischer, M., Kemper, P., 2000. Modeling and Analysis of a Freight Terminal with Stochastic Petri Nets. In *Proc. of 9th IFAC Int. Symp. Control in Transp. Systems*, Braunschweig, Germany, vol. 2, pp. 195-200.

Kelton, W.D., Sadowski, R.P., Sadowski, D.A., 1998. *Simulation with Arena*, McGraw Hill, New York.

Kozan, E., 2000. Optimising container transfers at multimodal terminals. *Mathematical and Computer Modelling*, Vol. 31, No. 10-12, pp. 235-243.

Legato, P., Mazza, R.M., Sept. 2001. Berth planning and resources optimisation at a container terminal via discrete event simulation. *European Journal of Operational Research*, Vol. 133, No. 3, pp. 537-547.

Liu, C.I., Ioannou, P.A., 2002. Petri Net Modeling and Analysis of Automated Container Terminal Using Automated Guided Vehicle Systems. *Transportation Research Record*, No. 1782, pp. 73-83.

Mastrolilli, M., Fornara, N., Gambardella, L.M., Rizzoli, A.E., Zaffalon, M., 1998. Simulation for policy evaluation, planning and decision support in an intermodal container terminal. In *Proc. Int. Workshop Modeling and Simulation within a Maritime Environment*, Riga, Latvia, 6-8 Sept. 1998, pp. 33-38.

Steenken, D., Voss, S., Stahlbock, R., 2004. Container terminal operation and operations research - a classification and literature review. *OR Spectrum*, 26, pp. 3-49.

Zeigler, B.P., Praehofer, H., Kim, T.G., 2000. *Theory of Modelling and Simulation*, Academic Press. New York, 2nd ed..

# MPC FOR SYSTEMS WITH VARIABLE TIME-DELAY
## *Robust Positive Invariant Set Approximations*

Sorin Olaru, Hichem Benlaoukli
*SUPELEC, Automatic Control Department, Gif-sur-Yvette, 91192, France*
*sorin.olaru@supelec.fr, hichem.belaoukli@supelec.fr*

Silviu-Iulian Niculescu
*LSS - SUPELEC, 3 rue Joliot Curie, F-91192 Gif-sur-Yvette, France*
*silviu.niculescu@lss.supelec.fr*

Keywords: Predictive control, Time-delay, Invariant sets.

Abstract: This paper deals with the control design for systems subject to constraints and affected by variable time-delay. The starting point is the construction of a predictive control law which guarantees the existence of a nonempty robust positive invariant (RPI) set with respect to the closed loop dynamics. In a second stage, an iterative algorithm is proposed in order to obtain an approximation of the maximal robust positive invariant set. The problem can be treated in the framework of piecewise affine systems due to the explicit formulations of the control law obtained via multiparametric programming.

## 1 INTRODUCTION

The delays (constant or time-varying, distributed or not) describe coupling between the dynamics, propagation and transport phenomena, heredity and competition in population dynamics. Various motivating examples and related discussions can be found in (Niculescu, 2001), (Michiels and Niculescu, 2007). There is an consensus in defining *delay* as a *critical parameter* in understanding dynamics behavior and/or improving (overall) system's performances. Independently of the mathematical problems related to the appropriate representation of such dynamics, the delay systems are known to rise challenging control problems due to the instabilities introduced by the deferred input actions. One of the natural ways to counteracting the effects of dead-time is to predict the system evolution but particular care has to be shown to the sensitivity of predictions for unstable models.

MPC - "Model Predictive Control" is a popular control technique based on the resolution of a finite-time optimal control problem over a receding horizon. Several strategies were proposed in order to reinforce the MPC stability (Maciejowski, 2002; Goodwin et al., 2004) having as main ingredients the terminal cost functions and the positive invariant terminal constraints (Mayne et al., 2000). Unfortunately, considering similar uncertainty interpretation in the context of time-varying delays lead to complex min-max optimization problems, difficult to handle on-line.

The present paper proposes an alternative issue for handling such a control problem. More precisely, we propose to use a simple MPC design constructed upon the nominal prediction. The resulting piecewise affine control law will transform the closed-loop dynamic in a piecewise affine system, the variable delay inducing in fact a model uncertainty. The existence of a non-empty positive invariant set can be guaranteed under mild conditions. Two problems related to the robustness of the designed control law will be dealt in detail:

- tuning the nominal MPC using inverse optimality;
- characterisation of the maximal robust positive invariant (MRPI) set.

It should be noted that the MRPI set may not be finitely determined and the second point will iteratively construct a dual expansive/contractive procedure for providing an inner approximation.

The paper is organized as follows: section 2 formulates the control problem and defines the models to be further used in the MPC design; section 3 deals with the construction of the explicit piecewise affine control law and section 4 details the approximation of the maximal robustly positive invariant set for the closed-loop system. Finally section 5 presents an example whereas section 6 draws the conclusions.

## 2 PROBLEM FORMULATION

Consider a linear continuous time system:

$$\dot{x} = A_c x(t) + B_c u(t-\tau) \tag{1}$$

affected by a variable time delay $\tau \in [0, \tau_{max}]$.

Note the discrete time instants $x_k = x(t_k) = x(kT_e)$ where $T_e$ is the sampling time. Consider:

$$\begin{array}{l} d = \lceil \bar{\tau}/T_e \rceil \\ \varepsilon = dT_e - \bar{\tau} \end{array} \tag{2}$$

where $\bar{\tau}$ is the "probable" value of the delay. The nominal discrete time LTI model is:

$$x_{k+1} = Ax_k + Bu_{k-d} - \bar{\Delta}(u_{k-d} - u_{k-d+1}) \tag{3}$$

The matrices $A, B, \bar{\Delta}$ are given by:

$$A = e^{A_c T_e} \tag{4}$$

$$B = \int_0^{T_e} e^{A_c(T_e-\theta)} B_c d\theta \tag{5}$$

$$\bar{\Delta} = \int_{-\varepsilon}^{0} e^{-A_c\theta} B_c d\theta \tag{6}$$

obtained by assuming that the control action $u$ is maintained constant $u(t) = u_k, \forall t \in [t_k, t_{k+1})$.

Despite this nominal model 3, in the general case, the variable time-delay implies a variable limit for the integration for $\varepsilon$. By considering $\Delta$ as a matrix affected by polytopic uncertainty (corresponding to the variation $0 \le \varepsilon \le T_e$) one can obtain the embedding:

$$\begin{array}{rcl} x_{k+1} & = & Ax_k + Bu_{k-v} - \Delta(u_{k-v} - u_{k-v+1}) \\ \Delta & \in & Co\{\Delta_0, \Delta_1, \ldots, \Delta_n\} \\ & & v \in \{0, 1, 2, \ldots, h\} \end{array} \tag{7}$$

where the maximum value of discrete delay is:

$$h = \left\lceil \frac{\tau_{max}}{T_e} \right\rceil \tag{8}$$

a set of $n+1$ vertices $\Delta_i$ are obtained using the Jordan form of $A_c$ (Olaru and Niculescu, 2008).

Using an extended state space representation based on the equation (3), one can obtain the nominal prediction model:

$$\xi_{k+1} = \bar{F}\xi_k + \bar{G}u_k \tag{9}$$

.

For the same state vector $\xi_k$, by using $A, B$ and the polytopic embedding for $\Delta$ with the extreme realizations $\Delta_i, i = \{0, \ldots, n\}$ in the extended state space, the polytopic model can be described:

$$\begin{array}{rcl} \xi_{k+1} & = & F\xi_k + Gu_k \\ (F,G) & \in & Co\{(F_1,G_1), \ldots, (F_s,G_s)\} \end{array} \tag{10}$$

with $s = nh + 1$.

The system evolution has to satisfy physical limitations leading to a set of linear inequalities:

$$C\xi_k \le W \tag{11}$$

The control objective is the regulation of the state $\xi_k$ to origin while satisfying the constraints using a receding horizon optimal control approach.

## 3 EXPLICIT CONTROL DESIGN: ROBUSTNESS ISSUE

### 3.1 Predictive Control

A standard MPC strategy, for the delay system considered here, will construct at each sampling instant $k$ the optimal control sequence:

$$\mathbf{k}_u^* = \{u_{k|k}, \ldots, u_{k+N-d-1|k}\} \tag{12}$$

with respect to a *performance index* which evaluates the system dynamics over a finite horizon $k+1, \ldots, k+N$. As a basic remark, the prediction horizon has to be larger than the delay $N \ge d$ in order to have an effective measure of its effect at the system output. Knowing that the prediction is constructed upon the nominal model but the real system may be affected by delays up to $h$ samples, it will be considered that $N \ge h$ in order to cope with all the possible variations.

The first component of $\mathbf{k}_u^*$ is effectively applied as control action to the system:

$$u_k = \mathbf{k}_u^*(1) = u_{k|k} \tag{13}$$

while the tail is discarded. Using the new measurements the optimisation procedure is restarted, thus obtaining a closed-loop control scheme.

The most popular performance index has a quadratic form and commensurate the state (tracking error) trajectory and the associated control effort. If the admissible trajectories are described by constraints as in (11), the MPC implementation passes by the resolution of the optimisation problem of the form:

$$\begin{array}{rl} \mathbf{k}_u^* = & \underset{\{u_{k|k}, \ldots, u_{k+N-d-1|k}\}}{arg\min} \Big\{ \xi_{k+N|k}{}^T \bar{P} \xi_{k+N|k} \\ & + \sum_{j=1}^{N} \xi_{k+j|k}{}^T \bar{Q} \xi_{k+j|k} + \sum_{j=0}^{N-d-1} u_{k+j|k}^T \bar{R} u_{k+j|k} \Big\} \end{array} \tag{14}$$

subject to:

$$\left\{ \begin{array}{l} \xi_{k+j+1|k} = \bar{F}\xi_{k+j|k} + \bar{G}u_{k+j|k} \\ C\xi_{k+j|k} \le W;\ j = 1, \ldots, N-1 \\ u_{k+i|k} = 0, \quad i = N-d, \ldots, N-1 \\ \xi_{k+N} \in X_N; \end{array} \right.$$

The construction of the predictive control law will be influenced by the choice of the prediction horizon $N$, the weighting factors on the state trajectory $\bar{Q} = \bar{Q}^T \succ 0$ and the control effort, $\bar{R} = \bar{R}^T \succ 0$. For the penalty on the terminal state the matrix $\bar{P}$ is usually constructed such that the prediction horizon to be extended to infinity by the introduction of the term

${\xi_{k+N|k}}^T \bar{P} \xi_{k+N|k}$ in (14). However $\xi_{k+N|k}$ has to satisfy some mild conditions materialized by the terminal constraint which force this prediction to reach the predefined invariant set $X_N$. The usual choice in this sense ((Gilbert and Tan, 1991)) is the maximal output admissible set $X_N = O_\infty$ constructed for the system (9) with the optimal control satisfying the discrete algebraic Riccati equation:

$$\begin{aligned} \bar{P} &= \bar{Q} + \bar{F}^T \bar{P} \bar{F} - \bar{K}^T \left(\bar{R} + \bar{G}^T \bar{P} G\right) \bar{K} \\ \bar{K} &= -\left(\bar{R} + \bar{G}^T \bar{P} G\right)^{-1} \bar{G}^T F \end{aligned} \tag{15}$$

This is the classical design for the MPC law. In the subsection 3.3, the choice of the performance index will be discussed (in particular the matrices $\bar{Q}, \bar{R}$ and indirectly $\bar{P}$) such that the resulting control law to present a certain degree of robustness with respect to the variable delay.

## 3.2 Multiparametric Programming

After expressing the predictions as functions of the current state and the future control action, the optimisation problem in (14) can be reformulated as a multiparametric quadratic problem ((Bemporad et al., 2002),(Goodwin et al., 2004),(Dua et al., 2007), (Olaru and Dumur, 2005))

$$\begin{aligned} &\mathbf{k}_u^*(\xi_k) = \arg\min_{\mathbf{k}_u} 0.5 \mathbf{k}_u^T H \mathbf{k}_u + \mathbf{k}_u^T G \xi_k \\ &\text{subject to :} \quad A_{in} \mathbf{k}_u \leq b_{in} + B_{in} \xi_k \end{aligned} \tag{16}$$

where the vector $\xi_k$ plays the role of parameter.

Further, explicit solutions for the MPC law can be obtained by retaining the first component of $\mathbf{k}_u^*(\xi_k)$, thus expressing the predictive control in terms of a piecewise affine feedback law:

$$u_k = K_i^{MPC} \xi + \kappa_i^{MPC}, \quad \text{with} \quad i \text{ s.t. } x \in D_i, \tag{17}$$

for $D_i$, polyhedral regions in $\Re^{n+hm}$.

**Remark 1.** *The prediction model is linear, the origin is a feasible point (in the most cases placed on the interior of the feasible domain) and thus represents an equilibrium point for the system (9). The problems (14), and further (16), are feasible and more then that, the associated optimum will be unconstrained.*

The consequence is that the affine control law corresponding to the region $D_{i_0}$ containing the origin ($0 \in D_{i_0}$) is in fact a linear feedback ($\kappa_{i_0}^{MPC} = 0$) and it corresponds to the unconstrained optimal control law ($K_{i_0}^{MPC} = K_{LG}$ if $\bar{P}$ is build upon (15)). If the constraints are symmetric, the region containing the origin will be the central region of the partition (the symmetry is inherited in the polyhedral decomposition of the state space).

## 3.3 Tuning MPC for Robustness

Consider an infinite-horizon min-max control problem for the polytopic system (10):

$$\min_K \quad \max_{F \in \Omega_\xi} \sum_{i=0}^{\infty} \xi_{k+i}^T Q \xi_{k+i} + u_{k+i}^T R u_{k+i} \tag{18}$$

$$u_k = K \xi_k \tag{19}$$

where $Q > 0$, $R > 0$ are suitable weighting matrices fixed *a priori* and $K$, the feedback gain playing in fact the role of the optimization argument.

Consider a quadratic function of the state

$$V(\xi) = \xi^T P \xi, \quad P > 0 \tag{20}$$

which represents an upper bound for $J_\infty$ if the following inequality is satisfied $\forall F \in \Omega_\xi$:

$$V(\xi_{k+i+1}) - V(\xi_{k+i}) \leq -[\xi_{k+i}^T Q \xi_{k+i} + u_{k+i}^T R u_{k+i}] \tag{21}$$

Rewriting this equation using (19) the following inequality is obtained:

$$\begin{aligned} \xi_{k+i}^T [&(F+GK)^T P (F+GK) \\ &- P + K^T R K + Q] \xi_{k+i} \leq 0 \end{aligned} \tag{22}$$

or equivalently:

$$(F+GK)^T P (F+GK) - P + K^T R K + Q \leq 0 \tag{23}$$

Using the ideas in (Boyd et al., 1994), by noting $P = GS^{-1}$ and $Y = KS$, for $S \geq I$, the following LMI can be constructed:

$$\begin{bmatrix} S & SF^T + Y^T G^T & SQ^{1/2} & Y^T R^{1/2} \\ FS + GY & S & 0 & 0 \\ Q^{1/2} S & 0 & GI & 0 \\ R^{1/2} Y & 0 & 0 & GI \end{bmatrix} \succ 0, \tag{24}$$

Using now the fact that $F \in \Omega_\xi$, a stabilizing control law is given by $K = YS^{-1}$ where $Y$, $S$ and the scalar $G$ are the solutions of the LMI problem (similar with the construction in (Kothare et al., 1996)):

$$\begin{aligned} &\min_{G,S,Y} G \\ &\begin{bmatrix} S & SF_i^T + Y^T G_i^T & SQ^{1/2} & Y^T R^{1/2} \\ F_i S + G_i Y & S & 0 & 0 \\ Q^{1/2} S & 0 & GI & 0 \\ R^{1/2} Y & 0 & 0 & GI \end{bmatrix} \succ 0, \\ &\text{for all} \quad i = 0, \ldots, s \\ &S \geq I \end{aligned} \tag{25}$$

**Remark 2.** *This LMI based procedure is used in (Kothare et al., 1996) to design a robust MPC law. The LMI in (25) is not depending on the measured state and thus the resulting control law is represented by a fixed feedback control gain.*

The resulting law $u_k = Kx_k$ represents a robust stabilizing control in the unconstrained case. In the sequel, the idea is to use this information when tuning the nominal MPC parameters in (14), namely $Q, R$ and $P$. We start with the remark that the MPC law is a piecewise affine function of the state and the central region (or the region containing the origin, if the constraints are not symmetric) is characterized by the unconstrained optimum for the chosen performance index in (14). Constructing this performance index such that the optimal solution corresponds to the LQ solution ($K = YS^{-1} \leftrightarrow K_{LQ}$) can be seen as an inverse optimality problem (Kalman, 1964).

Roughly speaking the tuning procedure is the following: given the matrices $\bar{F}, \bar{G}$ and $Y, S$ from (25), the matrices $\bar{Q} \geq 0$ and $\bar{R} > 0$ (and indirectly $\bar{P} \geq 0$) will be constructed such that the optimal solution to the unconstrained problem (14) to be:

$$\mathbf{k}_u^* = \begin{bmatrix} YS^{-1} \\ YS^{-1}(\bar{F} + \bar{G}YS^{-1}) \\ \vdots \\ YS^{-1}(\bar{F} + \bar{G}YS^{-1})^{N-1} \end{bmatrix} \xi_k \qquad (26)$$

The (not unique) pair $(\bar{Q}, \bar{R})$ has to satisfy:

$$\bar{Q} = \bar{P} - \bar{F}^T \bar{P} \bar{F} + \left\{YS^{-1}\right\}^T (\bar{R} + \bar{G}^T \bar{P} \bar{G}) YS^{-1} \qquad (27)$$

$$\bar{R}YS^{-1} + \bar{G}^T \bar{P} \bar{G} YS^{-1} + \bar{G}^T \bar{P} \bar{F} = 0 \qquad (28)$$

This problem can be solved in the general case by employing an LMI formulation (Larin, 2003):

$$\begin{gathered} \min \alpha \\ \bar{P} - \bar{F}^T \bar{P} \bar{F} + \left\{YS^{-1}\right\}^T (\bar{R} + \bar{G}^T \bar{P} \bar{G}) YS^{-1} \succ 0 \\ \begin{bmatrix} Z & \bar{R}YS^{-1} + B^T \bar{P} BYS^{-1} + B^T \bar{P} A \\ * & I \end{bmatrix} \succ 0 \\ Z \prec \alpha I, \quad \bar{P} \succ 0 \end{gathered} \qquad (29)$$

**Theorem 1.** *The nominal MPC control law, designed upon a performance index obtained by inverse optimality with respect to an unconstrained robust linear feedback, is robustly stabilizing the system (10) despite of constraints on a nondegenerate neighborhood of the origin $V$.*

**Proof:** The proof is constructive and follows the arguments described in this section. Using the LMI formulation (25), a robustly stabilizing control law is obtained for the unconstrained system (10) affected by uncertainty. The corresponding gain $K = YS^{-1}$ will be used together with the nominal model for the resolution of the LMI problem (29) which provides by inverse optimality the matrix $\bar{R}$. The matrix $\bar{Q}$ is obtained with a simple evaluation of (27) and the structure of the performance index in (14) is completed. The prediction horizon of the same performance index can be chosen according with the desired performances and complexity of the explicit solution. Independently of this choice, if the matrix $\bar{P}$ satisfies (15), then the nominal MPC leads to a piecewise affine control law and for the region $D_{i_0}$ with $0 \in Int(D_{i_0})$ the explicit control law will be

$$u_k = K_{i_0}^{MPC} \xi_k + \kappa_{i_0}^{MPC} = YS^{-1} \xi_k \qquad (30)$$

This region is polyhedral and the robust stabilizing properties are verified for an invariant subset with respect to the closed loop dynamics (10). If we consider the general form of the invariant set given by the level set:

$$E(\sigma) = \left\{\xi | \xi^T \bar{P} \xi \leq \sigma\right\} \qquad (31)$$

then one can find $\sigma > 0$ satisfying $V = E(\sigma) \subset D_{i_0}$. ■

# 4 RPI SET

The synthesis problem being solved, we dispose of a control law supposed to stabilize a time-varying delay system. The question is: *which is the maximal invariant set for the closed loop system?* An approximation can be obtained by constructing the maximal robust positive invariant set (MRPI) for a piecewise affine system (PWA) affected by uncertainty.

A PWA system is obtained from the embedding of the time-varying system in a linear model affected by polytopic uncertainty in closed loop with the piecewise affine control law:

$$\begin{aligned} \xi_{k+1} &= f_{PWA}(\xi_k) = (F + GK_i^{MPC})\xi_k + \kappa_i^{MPC} \\ &\qquad \text{for} \quad \xi_k \in D_i \\ (F, G) &\in Co\{(F_1, G_1), \ldots, (F_s, G_s)\} \end{aligned} \qquad (32)$$

where $D_i$ are the polytopic partition $D = \cup_i D_i$.

The dynamics related to an extreme realization of the PWA polytopic uncertainty will be described by:

$$\begin{gathered} \xi_{k+1} = f_{PWA_i}^j(\xi_k) = (F_j + G_j K_i^{MPC})\xi_k + \kappa_i^{MPC} \\ \text{for} \quad \xi_k \in D_i, j \in \{0, 1, \ldots, s\} \end{gathered} \qquad (33)$$

The description of the MRPI set for such a PWA system is not immediate, even for simple cases the finite determiness can not be guaranteed. Nevertheless, the fact that the partition of the state space is given by polyhedral regions will be used in the following section to build appropriate approximations.

In order to describe these geometrical constructions, the image and preimage operators over the sets $\Psi \in \Re^{n+hm}$ will be defined as:

$$\begin{aligned} Im_{f_{PWA}}(\Psi) = \bigcup_j \{ &\zeta \in \Re^{n+hm} | \exists \xi \in \Psi, \quad \text{s.t.} \\ &\zeta = (F_j + G_j K_i^{MPC})\xi + \kappa_i^{MPC} \quad \text{for} \quad \xi \in D_i \cap \Psi\} \end{aligned} \qquad (34)$$

$$PreIm_{f_{PWA}}(\Psi) = \bigcap_j \{\xi \in D | \exists \zeta \in \Psi, \quad \text{s.t.} \quad \zeta = (F_j + G_j K_i^{MPC})\xi + \kappa_i^{MPC} \text{ for } \quad \xi \in D_i\} \tag{35}$$

**Contractive Procedure:** The idea is to substract from the state partition $D = \cup_i D_i$ defining the PWA system, those regions for which one of the extreme dynamics will evolve outside $D$. This is an iterative procedure as long as after each iteration, the set $D$ is modified and thus the possible evolutions are to be rechecked.

The complexity of the procedure is given by the fact that the subtraction of convex set is not a closed operation. In short, if $D$ is convex, there is no guarantee that it will remain convex after an iteration of the contractive procedure. Indirectly this is acknowledging the fact that the MRPI set may not be convex.

**Algorithm 1:** Contractive Scheme

$V_0 = D$
$k = 0$
`while` *(precision condition)*
$V_{k+1} = PreIm_{f_{PWA}}(Im_{f_{PWA}}(V_k) \cap V_k)$
$k = k+1$

**Expansive Procedure:** In this case instead of excluding gradually those regions outside the MRPI set, we start with an RPI set and add those regions which evolve in one step inside the RPI set. Again the resulting set is RPI and is monotonically increasing (in the sense of inclusion) and is limited by MRPI.

An important advantage of the expansive procedure is that the intermediate results are robust positive invariant and thus can be considered as candidate approximations for the MRPI set.

**Algorithm 2:** Expansive Scheme

`find` $\sigma > 0$ s.t. $E(\sigma) \subset D_{i_0}$
$V_0 = E(\sigma)$
$k = 0$
`while` *(precision condition)*
$V_{k+1} = PreIm_{f_{PWA}}(Im_{f_{PWA}}(D) \cap V_k)$
$k = k+1$

Note the maximal robust positive invariant set $\Psi$ and the iterates obtained with the expansive and contractive procedure by $\Psi_i^e$ and $\Psi_i^c$ respectively.

Neither the expansive procedure $\Psi_i^e \subset \Psi$, nor the contractive procedure $\Psi_i^c \supset \Psi$ do not dispose of a measure of the convergence toward the MRPI set. However, by mixing the two relations we obtain an inner approximation for the MRPI set:

$$\Psi_i^e \subset \Psi \subset \Psi_i^c \tag{36}$$

Considering the Hausdorff metric over the class of polyhedra. The distance $d_H(\Psi_i^c, \Psi_i^e)$ can provide a measure of the MRPI approximation offered by $\Psi_i^e$ and thus a *precision condition*:

$$\Psi_i^e \subset \Psi \subset \Psi_i^c \subset \Psi_i^e \oplus \mathbf{B}_0(d_H(\Psi_i^c, \Psi_i^e)) \tag{37}$$

# 5 EXAMPLE

Consider the level control system as the one reported in (Furtmueller and del Re, 2006) with the bloc representation presented in figure 1. Beside the sensor and the actuator transfer functions we retrieve in this schema-block the variable time-delay; a nonlinear function $\Phi$ known and invertible and an integrator. The paper (Furtmueller and del Re, 2006) presented a method for the disturbance suppression, such that in the following we will consider the level control and replace the classical PI controller with a predictive controller and characterize the safety functioning region by the construction of the robust positive invariant region following the procedure presented in the previous sections.

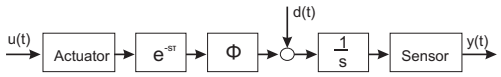



Figure 1: General scheme for the plant to be controlled.

The continuous time system to be controlled is a double integrator with variable-time delay and the discrete-time model is given by:

$$x_{k+1} = \begin{bmatrix} 1 & 0 \\ 0.1 & 1 \end{bmatrix} x_k + \begin{bmatrix} 0.1 \\ 0.05 \end{bmatrix} u_{k-i} - \Delta(u_{k-v} - u_{k-v+1}), \text{ with v} \in \{0,1,2\} \tag{38}$$

In the first instance the embedding of the uncertainty matrix $\Delta$ have to be obtained. Due to the fact that in the original representation, we deal with a 2-dimensional state vector $x_k$, the poytopic uncertainty will be:

$$\Delta \in Co\left\{ \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0.05 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0.013 \end{bmatrix} \right\} \tag{39}$$

In the extended state representation, a robustly stabilizing feedback gain is obtained for the unconstrained case by solving the LMI problem (25):

$$K = [\ -1.3188 \quad -0.5408 \quad -0.1292 \quad -0.0157 \quad -0.1511\ ] \tag{40}$$

The inverse optimality problem leads after solving (29) to the tuning of the nominal MPC law with the weighting matrices $P, Q, R$. By imposing a set of constraints on the input and the state:

$$-0.1 \leq u_k \leq 0.1$$
$$\begin{bmatrix} -2 \\ -2 \end{bmatrix} \leq x_k \leq \begin{bmatrix} 2 \\ 2 \end{bmatrix} \tag{41}$$

and taking into account that the maximal delay is 3 sampling instants we choose a prediction horizon $N =$

5 in order to maintain a low complexity of the explicit solution (47 regions in the state space partition, see figure 2).

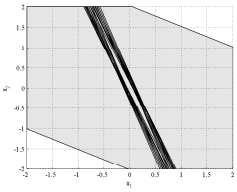


Figure 2: Projection of the explicit solution's partition on the first two components of the extended state space.

The polytopic model in the extended state representation which embeds (using 7 extreme realizations) the time-varying delay system will allow the use of the contractive procedure for the approximation of the maximal invariant set. In figure (3) cuttings through the approximation obtained after 5 iterations is presented.

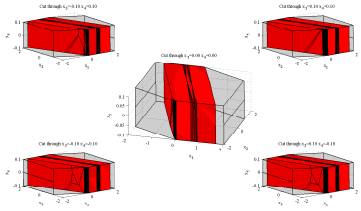


Figure 3: The explicit solution's partition and the approximation of the MRPI set.

Finally in figure (4-5) a time domain simulation with varying delay is presented (starting from the state $(0;-2)$), proving the versatility of the proposed control technique.

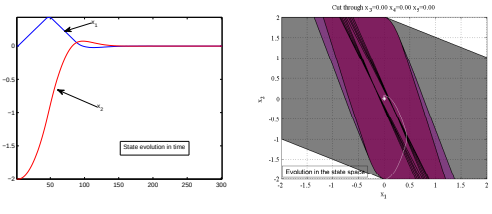



Figure 4: The time evolution of the state components.

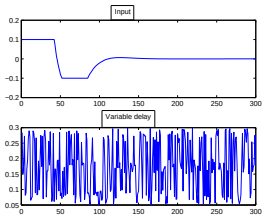



Figure 5: The control signal and the variation of the delay in time.

# 6 CONCLUSIONS

A model predictive control law was designed to deal with time-varying delay systems. The constraints are handled from the design stage and the iterative approximation of the maximal positive invariant set offers information about the region of the state space where the control policy is viable.

# REFERENCES

Bemporad, A., Morari, M., Dua, V., and Pistikopoulos, E. (2002). The explicit linear quadratic regulator for constrained systems. *Automatica*, 38:3–20.

Boyd, S., Ghaoui, L. E., Feron, E., and Balakrishnan, V. (1994). *Linear Matrix Inequalities in System and Control Theory*. SIAM, Philadelphia, USA.

Dua, V., Pistikopoulos, E. N., and Georgiadis, M. C. (2007). *Multi-Parametric Model-Based Control: Theory and Applications*. Wiley-VCH Verlag, Germany.

Furtmueller, C. and del Re, L. (2006). Disturbance suppression for an industrial level control system with uncertain input delay and uncertain gain. In *Proc. of the IEEE Conf. on Control Applications*, pages 3206–3211.

Gilbert, E. and Tan, K. (1991). Linear systems with state and control constraints: The theory and application of maximal output admissible sets. *IEEE Transactions on Automatic Control*, 36:1008–1020.

Goodwin, G., Seron, M., and Dona, J. D. (2004). *Constrained Control and Estimation*. Springer, Berlin.

Kalman, R. E. (1964). When is a linear control system optimal? *Trans. ASME, Journal of Basic Engineering, Series D*, 86:81–90.

Kothare, M., Balakrishnan, V., and Morari, M. (1996). Robust constrained model predictive control using linear matrix inequalities. *Automatica*, 32:1361–1379.

Larin, V. (2003). About the inverse problem of optimal control. *Journal of Applied and Computational Mathematics*, 2:90–97.

Maciejowski, J. (2002). *Predictive Control with Constraints*. Prentice Hall, England.

Mayne, D., Rawlings, J., Rao, C., and Scokaert, P. (2000). Constrained model predictive control: Stability and optimality. *Automatica*, 36:789–814.

Michiels, W. and Niculescu, S.-I. (2007). *Stability and stabilization of time-delay systems. An eigenvalue based approach*. SIAM, Philadelphia, USA.

Niculescu, S.-I. (2001). *Delay effects on stability. A robust control approach*. Springer, Heidelberg.

Olaru, S. and Dumur, D. (2005). Avoiding constraints redundancy in predictive control optimization routines. *IEEE Transactions on Automatic Control*, 50(9):1459–1466.

Olaru, S. and Niculescu, S.-I. (2008). Predictive control for linear systems with delayed input subject to constraints. In *Proceedings of the IFAC World Congress*.

# SYNTHESIS OF VELOCITY REFERENCE CAM FUNCTIONS FOR SMOOTH OPERATION OF HIGH SPEED MECHANISMS

Robert M. C. Rayner and M. Necip Sahinkaya
*Department of Mechanical Engineering, University of Bath, Claverton Down, Bath BA2 7AY, U.K.*
*r.m.rayner@bath.ac.uk, ensmns@bath.ac.uk*

Keywords: Cam function, mechanisms, identification, control, simulation.

Abstract: The purpose of the paper is to improve the dynamic performance of a mechanism used in a packaging machine in order to run the system at higher speeds with lower vibration and noise levels. A method of synthesising a velocity demand signal as a function of crank position (i.e. cam function) is demonstrated for a prototype mechanism and drive system. The method aims to minimise the peak to peak actuation torque requirements in order to minimise the vibration of the mechanism. First of all, experimental results are utilised to identify the drive system parameters. A dynamic simulation package is used to model the nonlinear dynamics of the mechanism. The model based synthesis of velocity reference cam functions is performed at increasing mechanism actuation speeds. The performance of the system using the proposed velocity demand cam function is compared with the conventional constant speed reference case at different running speeds.

## 1 INTRODUCTION

The dynamic performance requirements of modern machinery are constantly increasing in terms of operation speed and motional accuracy. To remain competitive, mechanisms need to run at ever higher speeds, with greater reliability and be manufactured at lower cost. To achieve this, machines use a combination of electrical control systems, servo systems and mechanisms to generate truly mechatronic solutions. At the core of most packaging machines are multi-linkage mechanisms, which interact with packaging materials and products. These mechanisms have highly nonlinear dynamic properties and introduce vibrations at high operating speeds.

Much work has been documented on optimum balancing of mechanisms in order to reduce the vibrations at high operating speeds, such as (Kochev, 2000; Lee and Cheng, 1984; Alici and Shirinzadeh, 2006) and others. This method involves the adding of balancing masses to the mechanism, which increases the weight and may not always be physically achievable due to factors such as space restrictions. Alternatively, a mechanism can be re-designed or re-synthesised by considering kinematic and dynamic cost functions (Conte et al., 1975; Kochev, 2000). Due to the large number of parameters, conventional optimisation techniques struggle. Many researchers tried to formulate new optimisation techniques, such as genetic algorithms (Connor et al., 1998; Cabrera et al., 2002; Laribi et al., 2004; Saxena, 2005), differential evolution (Price and Storn, 2006), artificial immune searching (Liu and Xiao, 2005), geometric centroid of precision positions technique (Shiakolas et al., 2005), and the time varying dimensions method (Hansen, 2002).

It has been stated in (Yuan and Rastegar, 2004) that vibrations experienced during high speed actuation are caused by harmonic content in the output motion. It has also been argued in (Rastegar and Yuan, 2002) that the amount of harmonic content present in the motion increases with the magnitude of the peak to peak torque required to generate the motion. Recently, an iterative method of synthesising a velocity command cam function has been introduced to reduce the peak to peak actuation torque (Sahinkaya et al., 2007). This method relies on the development of a computer model of the system. This model is based on experimental results and a simulation of the nonlinear dynamics of the mechanism. This paper extends the aforementioned work by synthesising optimised cam functions (i.e. velocity profile as a function of crank angle) to achieve higher output speeds than that discussed previously, and that used for the purpose of system identification.The use of shaped cam functions has demonstrated significant benefits in terms of reduced peak to peak torque requirements.This is a software based command shaping technique. No redesigning or re-synthesis of the mechanism is necessary.

## 2 SYSTEM DESCRIPTION AND MODELLING

The block diagram of the prototype system considered in this study is shown in Figure 1. The mechanism is a 6-bar mechanism called the *woodpecker* mechanism. The purpose of the mechanism is to push thin products into packaging held in a neighbouring hopper. The mechanism is driven by an Allen Bradley MPL 540K-MJ22AA servo controlled with an Allen Bradley Kinetix6000 drive unit via a belt drive with a 3:1 gear ratio. The drive unit is fundamentally a PI controller. The user can configure the drive unit and monitor the system in real-time using RSLogix5000 control software. The servomotor and the drive system is assumed to be a first order lag with a time constant of $\tau$ and a gain $K_m$.

$$G_m = \frac{K_m}{1+\tau s} \tag{1}$$

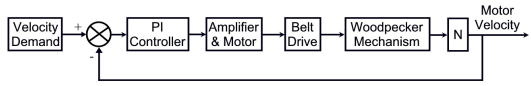



Figure 1: Block diagram of the overall system.

Before modelling the nonlinear dynamics of the woodpecker mechanism, experiments were carried out to identify the drive system parameters. It was possible to log the velocity demand, velocity output, acceleration output, position output, and motor torque. A step input velocity signal was used with different values of $K_P$ and constant velocity demand. The integral gain $K_I$ was set to zero during the identification process. To estimate the effective friction coefficient acting on the motor shaft, the steady state response over a single crank cycle was considered. By using the approximate constant speed section of the cycle, the friction coefficient $b$ can be estimated from the following transfer function between the motor torque (the output of "Amplifier & Motor" block in Fig. 1) and the motor velocity:

$$\frac{T_{m,ss}}{\dot{\theta}_{m,ss}} = \frac{N^2}{b} \tag{2}$$

where $T_{m,ss}$ and $\dot{\theta}_{m,ss}$ are the average motor torque and motor speed respectively along the constant speed section of the steady state cycle, and $N$ is the gear ratio, i.e. $N = 3$. This gave a representative friction coefficient of $b = 0.255$. In order to identify the motor/amplifier gain $K_m$ and the effective rotor inertia, tests were repeated by replacing the woodpecker mechanism with a disk of known inertia. Thus the steady state gain of the closed loop system can be written as:

$$\frac{\dot{\theta}_{m,ss}}{\dot{\theta}_{m,ref}} = \frac{N^2 K_m K_p}{b + N^2 K_m K_p} \tag{3}$$

By analysing the steady state system response, the motor/amplifier gain was estimated as $K_m = 0.0883$. Transient motor torque and motor velocity data were used to determine the effective inertia of motor shaft and associated pulleys acting on the crank shaft. Thus $I_m = 0.0071$ kgm$^2$. Observation of the transient motor torque and the error signal suggested that the time constant $\tau$ can be taken as zero. This may be due to a high-gain internal current feedback in the motor drive circuit.

The dynamic model for the woodpecker mechanism was built in Simulink by using *Dysim* (Hazlerigg and Sahinkaya, 1984; Sahinkaya, 2004) simulation package. The mechanism consists of 6 links as shown in Figure 2. The physical data is given in Table 1. A CAD model of the mechanism was used to obtain these mass and inertia values and the position of the centre of gravity of each link and local coordinates of the connection points. The model was then tested using various demand speed and controller parameter combinations. For example, Fig. 3 shows the experimental and simulation results for a constant speed reference of 300 rpm with $K_P = 20$ and $K_I = 0$.

The results showed an excellent match between the experimental and simulated responses for all the tests conducted with the prototype system. The high

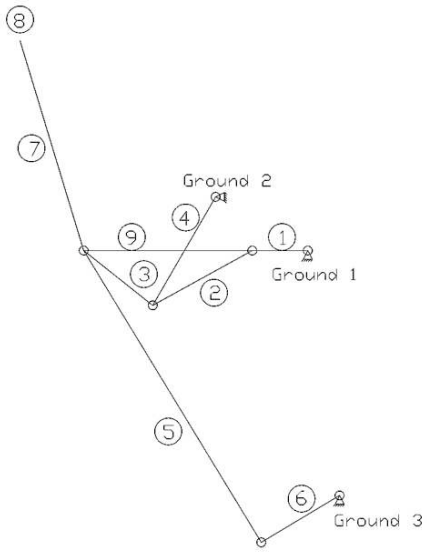



Figure 2: A schematic view of the woodpecker mechanism.

Table 1: Data for the Woodpecker mechanism.

| Name | No | Length (mm) | Mass (kg) | Inertia ($N{\cdot}mm^2$) |
|---|---|---|---|---|
| Crank | 1 | 62 | 0.927 | 901 |
| Connector | 2<br>3<br>9 | 127<br>103<br>188.3 | 0.310 | 1420 |
| Upper pivot | 4 | 144 | 0.414 | 1310 |
| End-effector | 8 | | 0.174 | 380 |
| Spine | 5<br>7 | 348.86<br>245.19 | 0.482 | 10550 |
| Lower pivot | 6 | 102 | 0.123 | 290 |

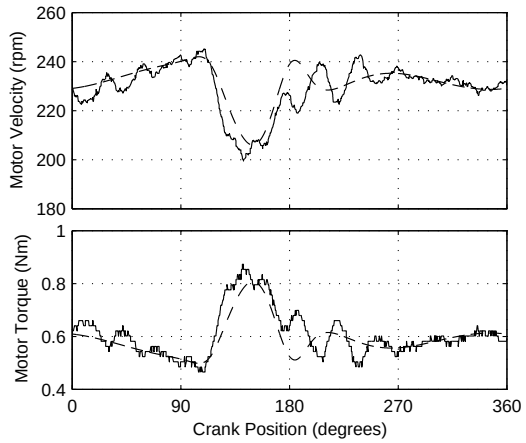



Figure 3: Simulated (dashed) and experimental (solid) steady state responses over one crank cycle.

frequency oscillations (of approximately at 15 Hz.) in the measured response were due to the belt dynamics, which were not included in the analysis. Particularly encouraging was the reproduction in the simulated results of the velocity trough and corresponding torque peak resulting from the nonlinear nature of the mechanism dynamics. The orbit of the end effector is shown in Figure 4. Normalised crank positions (unity normalised crank position corresponds to $360^o$ crank angle) are shown on the orbit. The critical portion of the path is between normalised crank positions of 0.6 and 1.0, where the end-effector interacts with the product and product feeding mechanism.

## 3 OPTIMUM CAM PROFILE

The experimental results highlighted a potential problem when running the system at higher speeds. Of particular concern was the torque spike and trough on the return part of the end-effector orbit between crank positions 90 and 180 degrees. It has been shown elsewhere (Yuan and Rastegar, 2004) that harmonic content in the output motion induces vibra-

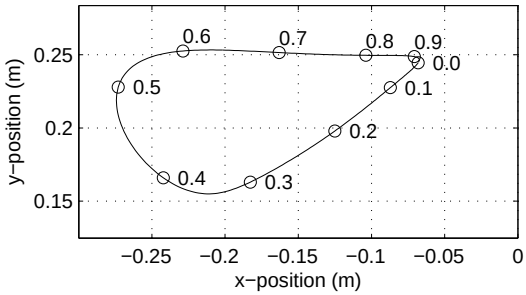



Figure 4: Orbit of the end-effector.

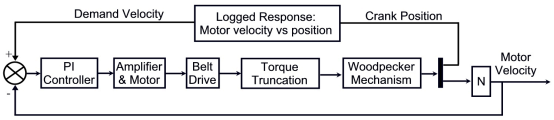



Figure 5: The process of optimising a velocity cam function.

tions and that the amount of harmonic content increases with the peak-to-peak magnitude of the actuation torque (Rastegar and Yuan, 2002). Therefore, the focus of the paper is to reduce the peak-to-peak drive torque through shaping the speed reference signal as a function of crank angle. The machanisms will not be re-synthesise nor rebuilt. The method suggested in (Sahinkaya et al., 2007) utilises the model of the drive system estimated from experimental results. The procedure can be summarised as follows:

(a) Run the simulation of the overall system for a constant speed reference signal, and determine a narrow torque band from the steady state torque signal covering the approximate constant torque region.

(b) Run the simulation again with the same constant speed reference signal, but with saturation limits imposed on the drive torque. These limits are determined in (a). Then record the steady state output speed response over a single cycle of crank logged against crank position.

(c) Use the periodic output speed recorded in (b) as a velocity cam function and run the simulation without saturation limits to assess the performance of the system with this velocity cam function.

This process is shown in Figure 5 as a block diagram.

The above process is applied to the model of the prototype system to assess the benefit of the velocity cam function when the average running speed of the mechanism is increased from to 100 rpm to 600 rpm. Due to the 3:1 gear ratio, this corresponds to

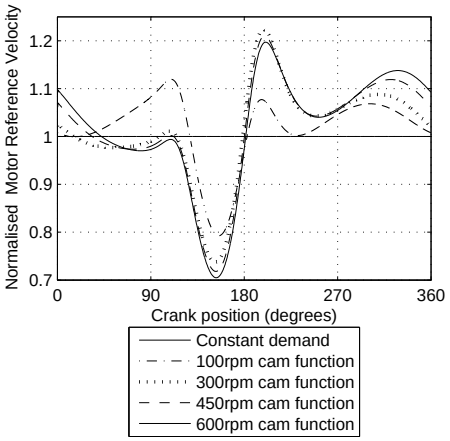



Figure 6: Velocity reference cam functions at different average crank speeds.

motor speeds from 300 rpm to 1800 rpm. The controller parameters are set to $K_P = 20$ and $K_I = 400$. Figure 6 shows the synthesised cam functions at 300, 900, 1350, and 1800 rpm of the motor speed. For ease of comparison, the velocity reference signals are normalised by their corresponding constant speed values. Note that in each case the achieved average cyclic velocity of the mechanism corresponded closely to the demand velocity.

Figure 7 shows the corresponding steady state crank velocity output over a single crank cycle. Especially at higher speeds, the change in system response is minimal compared with the constant speed reference signal cases. Despite small variations in the output velocity profile, the use of the optimised cam function has significant benefits, greatly reducing the peak to peak drive torque requirements as shown in Figure 8. The benefit of the optimised cam function can be better appreciated from Figure 9, where the reduction in peak to peak torque variations are 99%, 80%, 78%, and 78% for the crank speeds of 100, 300, 450, and 600 rpm respectively.

Although it is not included in the optimisation process, the optimised velocity cam functions also reduce the maximum drive motor power requirements compared with the constant velocity signal case as shown in Figure 10.

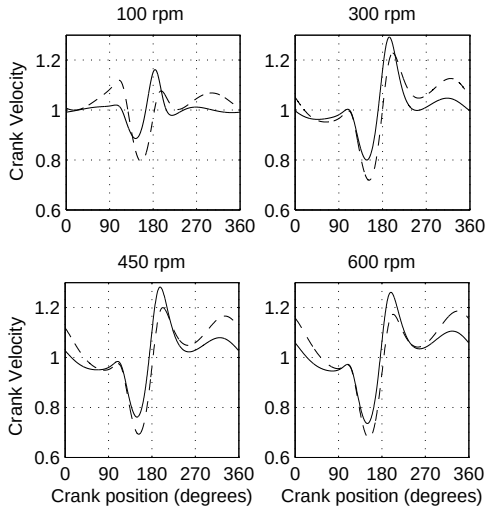



Figure 7: Normalised crank velocity output at different crank speeds (solid: constant reference, dashed: shaped reference).

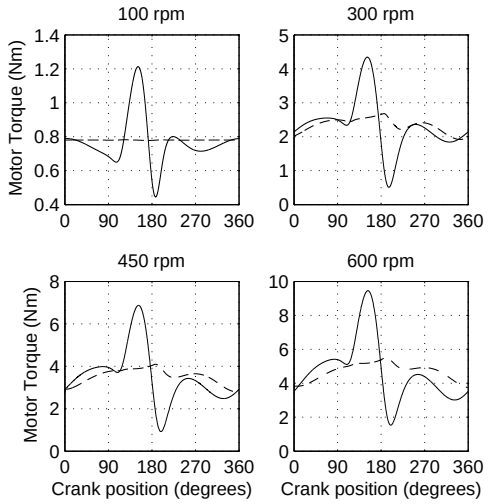



Figure 8: Drive torque at different crank speeds (solid: constant reference, dashed: shaped reference).

## 4 CONCLUSIONS

This paper demonstrates the benefit of using a velocity cam function as a velocity demand signal to reduce the peak to peak actuation torque of a servo driven mechanism. The drive system parameters were identified using experimental data, and then combined with a nonlinear dynamic model of the mechanism. The identification was carried out at a crank speed of 100 rpm. The accuracy of the computer model has been verified using experimental results. Utilising the computer model, a three-stage synthesis of the velocity demand cam functions has been performed at much higher operational speeds up to 600 rpm. The results show that a reduction in peak to peak actuation torque of as much as 80% can be achieved at high speeds without significantly affecting the speed response of the system. Although no effort has been made to minimise the energy consumption, sizable reductions in the maximum motor power requirements have also been predicted.

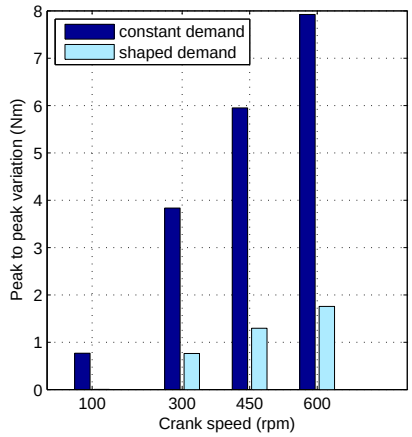



Figure 9: Peak to peak drive torque variations at different crank speeds.

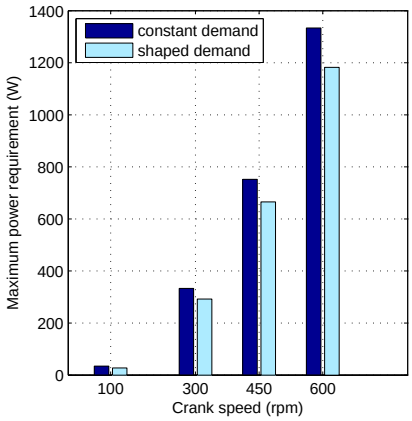



Figure 10: Maximum drive power requirements at different crank speeds.

## ACKNOWLEDGEMENTS

The authors acknowledge the support of the Engineering and Physical Sciences Research Council (EPSRC) of the U.K. and the industrial partner ITCM, Coventry, UK through the EPSRC Industrial Case Studentship Award Voucher No: 05002188.

## REFERENCES

Alici, G. and Shirinzadeh, B. (2006). Optimum dynamic balancing of planar parallel manipulators based on sensitivity analysis. *Mechanism and Machine Theory*, 41(12):1520–1532.

Cabrera, J. A., Simon, A., and Prado, M. (2002). Optimal synthesis of mechanisms with genetic algorithms. *Mechanism and Machine Theory*, 37(10):1165–1177.

Connor, A. M., Douglas, S. S., and Gilmartin, M. J. (1998). The use of harmonic information in the optimal synthesis of mechanisms. *Journal of Engineering Design*, 9(3):239–249.

Conte, F. L., George, G. R., Mayne, R. W., and Sadler, J. P. (1975). Optimum mechanism design combining kinematic and dynamic-force considerations. *Journal of Engineering for Industry-Transactions of the Asme*, 97(2):662–670.

Hansen, J. M. (2002). Synthesis of mechanisms using time-varying dimensions. *Multibody System Dynamics*, 7(1):127–144.

Hazlerigg, A. D. G. and Sahinkaya, M. N. (1984). Computer aided design of non-linear systems. In *ACC 84 Conference*, pages 1498–1503.

Kochev, I. S. (2000). General theory of complete shaking moment balancing of planar linkages: a critical review. *Mechanism and Machine Theory*, 35(11):1501–1514.

Laribi, M. A., Mlika, A., Romdhane, L., and Zeghloul, S. (2004). A combined genetic algorithm-fuzzy logic method (ga-fl) in mechanisms synthesis. *Mechanism and Machine Theory*, 39(7):717–735.

Lee, T. W. and Cheng, C. (1984). Optimum balancing of combined shaking force, shaking moment, and torque fluctuations in high-speed linkages. *Journal of Mechanisms Transmissions and Automation in Design-Transactions of the Asme*, 106(2):242–251.

Liu, Y. and Xiao, R. B. (2005). Optimal synthesis of mechanisms for path generation using refined numerical representation based model and ais based searching method. *Journal of Mechanical Design*, 127(4):688–691.

Price, K. and Storn, R. (2006). Differential evolution (DE), http://www.icsi.berkeley.edu/ storn/code.html. Electronic Citation.

Rastegar, T. and Yuan, L. (2002). A systematic method for kinematics synthesis of high-speed mechanisms with optimally integrated smart materials. *Journal of Mechanical Design*, 124(1):14–20.

Sahinkaya, M. N. (2004). Inverse dynamic analysis of multiphysics systems. *Proceedings of the Institution of Mechanical Engineers Part I- Journal of Systems and Control Engineering*, 218(I1):13–26.

Sahinkaya, M. N., Rayner, R. M. C., Vernon, G., Shirley, G., and Aggarwal, R. K. (2007). Synthesis of demand signals for high speed operation of a packaging mechanism. In *ASME International Design Engineering Technical Conferences & Computers and Information in Engineering Conference (IDETC)*.

Saxena, A. (2005). Synthesis of compliant mechanisms for path generation using genetic algorithm. *Journal of Mechanical Design*, 127(4):745–752.

Shiakolas, P. S., Koladiya, D., and Kebrle, J. (2005). On the optimum synthesis of six-bar linkages using differential evolution and the geometric centroid of precision positions technique. *Mechanism and Machine Theory*, 40(3):319–335.

Yuan, L. F. and Rastegar, J. S. (2004). Kinematics synthesis of linkage mechanisms with cam integrated joints for controlled harmonic content of the output motion. *Journal of Mechanical Design*, 126(1):135–142.

# EXPERIMENTAL OPEN-LOOP AND CLOSED-LOOP IDENTIFICATION OF A MULTI-MASS ELECTROMECHANICAL SERVO SYSTEM

Usama Abou-Zayed, Mahmoud Ashry and Tim Breikin
*Control Systems Centre, The University of Manchester, PO BOX 88, M60 1QD, U.K.*
*usama.abou-zayed@postgrad.manchester.ac.uk*

Keywords: System identification, black-box model, recursive least square algorithm, local optimal controller, and multi-mass servo systems.

Abstract: The procedure of system identification of multi-mass servo system using different methods is described in this paper. Different black-box models are identified. Previous experimental results show that a model consisting of three-masses connected by springs and dampers gives an acceptable description of the dynamics of the servo system. However, this work shows that a lower order black-box model, identified using off-line or on-line experiments, gives better fit. The purpose of this contribution is to present experimental identification of a multi-mass servo system using different algorithms.

## 1 INTRODUCTION

An important step in designing a control system is proper modeling of the system to be controlled. An exact system model should produce output responses similar to those of the actual system. The complexity of most physical systems makes the development of exact models infeasible. Therefore, in order to design a controller that is reliable and easy to understand in practice, simplified system models should be obtained around operating points and\or model order reduction (Ziaei, 2000).

System identification is an established modeling tool in engineering and numerous successful applications have been reported. The theory is well developed (Ljung, 1999; Soderstrom, 1989), and there are powerful software tools available, e.g., the System Identification Toolbox (SIT) (Ljung, 1997).

Different physical models of electromechanical servo systems based on multi-mass representation were discussed in (Abou-Zayed, 2008). Using grey-box off-line identification, inertial parameters and parameters describing flexibilities were identified. The physical parameters estimates showed no variations in the mechanical parameters, and acceptable variations in the electrical parameters. Experimental results in (Abou-Zayed, 2008) show that a model consisting of three masses connected by springs and dampers gives an acceptable description of the dynamics of the servo system. However, this model is a six-order state-space model.

The objective of this paper is to present our recent experimental studies on black-box open-loop and closed-loop identification of a three-mass electromechanical system. The closed-loop tests are performed using a local-optimal controller.

The paper is organized as follows. In section 2, the servo system is described briefly. In Section 3, the results of black-box off-line identification are presented. On-line open-loop and closed-loop identification of the studied system is discussed in section 4. Finally, Section 5 contains some conclusions.

## 2 EXPERIMENTAL SETUP

A view from the experimental setup is shown in Fig.1. The DC servo mechanism setup to be studied operates at ±10V input voltage with a permissible output motor shaft speed of 2200 r.p.m. The shaft is connected to an inertial load through a coupling gear with ratio (r=1/30).The load shaft carries an absolute position sensor with linear range ±10V. A personal computer PC (Pentium III, 700 MHz, 256 MB RAM), running the MATLAB software, is

connected to the servo system setup through a data acquisition card. This PC is used as a signal generator for the servo system input. It also used as a data logger to store the relevant system parameters at fixed sample time. The third function is a digital controller for closed-loop identification purposes.

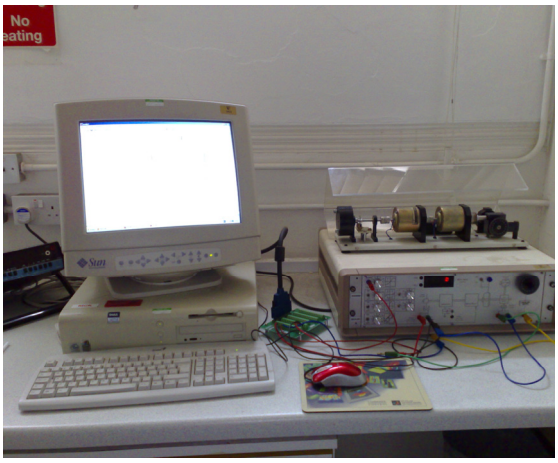


Figure 1: Experimental setup.

The DC servo system setup, shown in Fig.1 can be viewed as single input single output (SISO) for the present case, where the motor armature voltage $v_a$ is the input, while the output is the angular position of the load $\theta_L$. Since the measurement noise is fairly small, a reasonable estimate of the load angular speed $\omega_L$ is obtained for the identification purpose. Therefore, the load angular speed will be used as the output signal.

# 3 OFF-LINE IDENTIFICATION OF SERVO SYSTEM

This section presents the results of the study and realization of the off-line identification of servo systems for different types of models with different excitations. First, some dynamical properties of the system are obtained using the process reaction curve method. Then, black-box models, describing the system, are identified.

## 3.1 Process Reaction Curve Identification

It is one of the widely used approaches to predetermine the dynamic behaviour of a system before performing the data collection for system identification. An input step signal change is applied to the system, and the output response is measured. Rise time, settling time, bandwidth, time constant, time delay, and type of response can be determined using the Process reaction curve (Ziegler, 1942).

System step response is shown in Fig.2. The sample time chosen for this step test is 0.01sec to observe the system dynamical behaviour. An 8V input voltage (dashed line) is applied to the system. The output response (solid line) acts like a first order plus time delay system with average steady state output 7.18V, rise time about 1.4sec, and bandwidth around 1.6rad/s. Using the process reaction curve method the system can be modeled as in the classical case:

$$G(s) = \frac{Ke^{-T_d s}}{\tau s + 1} \qquad (1)$$

where $K$ is the steady state gain ($K$ = 0.898), $T_d$ is the time delay ($T_d$ = 10 ms), and $\tau$ is the time constant ($\tau$ = 0.73sec).

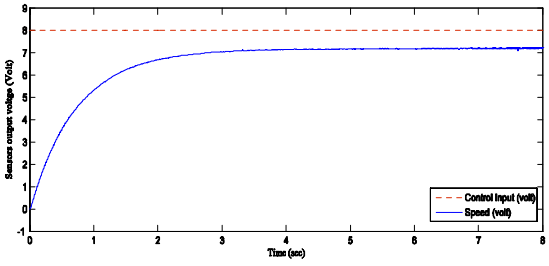



Figure 2: Output response of a step input change.

## 3.2 Experiment Design

The results of the identification experiments reported here are based on two data sets where the excitation signal has different character:

**Set 1:** A sum of 16 sinusoids with amplitude 1.8 and equidistant spacing in frequency, between 0.1 and 6.1 rad/sec. The resulting crest factor (Ljung, 1999) is 1.8 due to the Schroeder phase choice (Schroeder, 1970). The time response and power spectrum of the set are shown in Fig.3.

**Set 2:** A linear swept-frequency sinusoidal signal with amplitude 9 and time-varying frequency over a certain band ranges from 0.1 to 6.1rad/sec over a certain time period 100sec. The resulting input signal has a crest factor 1.42, and shown in Fig 4.

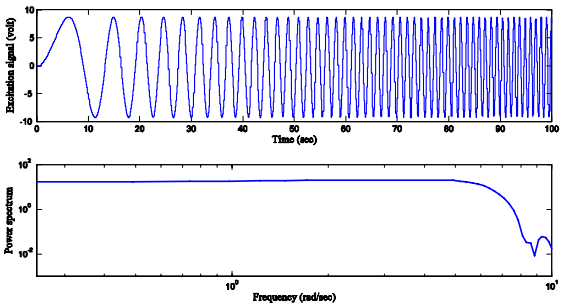



Figure 3: Multi-sine signal time response and power spectrum.

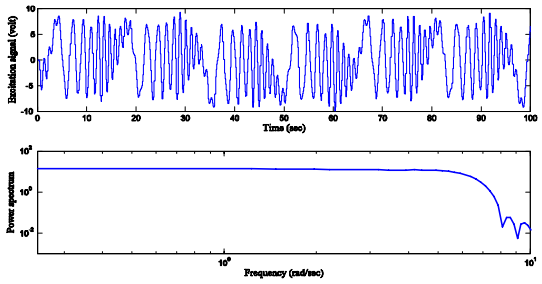



Figure 4: Chirp sine signal time response and power spectrum.

### 3.3 Black-box Transfer Function Model Identification

The starting point is the general linear model structure (Ljung, 1999),

$$y(t) = G(q,\theta)u(t) + H(q,\theta)e(t) \tag{2}$$

where q denotes the shift operator.

Two different model structures will be studied, and these are the ARX structure, defined by:

$$G(q,\theta) = B(q,\theta)/A(q,\theta), \quad H(q,\theta) = 1/A(q,\theta) \tag{3}$$

and the OE structure, where:

$$G(q,\theta) = B(q,\theta)/F(q,\theta), \quad H(q,\theta) = 1 \tag{4}$$

For the two model structures mentioned above, the estimation of the model parameters will be carried out generally using prediction error method (PEM). The identification experiments are carried out using the SIT (Ljung, 1997).

Tables 1, and 2 show the results of the estimated models using data set1 for ARX and OE model structures respectively. Both data sets show nearly similar estimates. The notation (ModelStructure pzd) denotes the p$^{th}$ order model with 'z' zeroes and delay'd'. The comparison is carried out using two different quantities. The first is MSE as:

$$MSE = \frac{1}{N}\sum_{t=1}^{N}(y(t)-\hat{y}(t))^2 \tag{5}$$

The second is the FIT:

$$\text{FIT} = \left(1-\left(\sqrt{\sum_{t=1}^{N}(y(t)-\hat{y}(t))^2 \Big/ \sum_{t=1}^{N}(y(t)-\bar{y})^2}\right)\right)\times 100\% \tag{6}$$

$\bar{y}$ is the mean value of the measured output.

Using k-step ahead predictors $\hat{y}_k = \hat{y}(t \mid t-k;\theta)$.The two extreme predictors is defined as:

$$\hat{y}_1(t) = H^{-1}(q)G(q)u(t) + \left[1-H^{-1}(q)\right]y(t) \tag{7}$$

$$\hat{y}_\infty(t) = G(q)u(t) \tag{8}$$

Table 1: Comparison of black-box ARX models.

| Model | MSE×10$^{-3}$ | fit (cross validation ) % | |
|---|---|---|---|
| | | $k$=1 | $k$=∞ |
| ARX 211 | 5.95 | 96.01 | 85.80 |
| ARX 311 | 2.29 | 97.72 | 80.25 |
| ARX 411 | 0.76 | 98.44 | 84.28 |
| ARX 511 | 0.58 | 98.76 | 83.00 |
| ARX 611 | 0.35 | 99.03 | 82.50 |

Table 2: Comparison of black-box OE models.

| Model | MSE×10$^{-3}$ | fit % |
|---|---|---|
| OE 211 | 39.90 | 85.53 |
| OE 311 | 38.20 | 84.99 |
| OE 321 | 38.30 | 84.98 |
| OE 421 | 37.60 | 85.66 |
| OE 611 | 37.70 | 85.72 |
| OE 621 | 34.70 | 86.17 |

It is clear that for OE models, there is no difference between both predictors. Otherwise, there is considerable difference between them. The one-step ahead predictor can give fits that "look good," even though the model may be bad. Therefore, the simulation fit can be used for invalidating the bad models.

# 4 ON-LINE IDENTIFICATION OF SERVO SYSTEM

## 4.1 Open-loop System Identification

Experiments are performed to find the discrete-time model that can best represent the system using RLS method. Let the system model is given in the form:

$$A(z^{-1})y(t) = B(z^{-1})u(t-1) \tag{9}$$

where $z^{-1}$ is the back shift operator, and

$$\begin{aligned} A(z^{-1}) &= 1 + a_1 z^{-1} + a_2 z^{-2} + \cdots\cdots + a_{n_a} z^{-n_a} \\ B(z^{-1}) &= b_o + b_1 z^{-1} + b_2 z^{-2} + \cdots\cdots + b_{n_b} z^{-n_b} \end{aligned} \tag{10}$$

A model of the system in (9) can be presented in the form of

$$y(t) = \varphi^T(t)\theta \tag{11}$$

where $\theta$ is a vector of unknown parameters defined by:

$$\theta = \left[a_1, \cdots\cdots, a_{n_a}, b_o, \cdots\cdots, b_{n_b}\right]^T \tag{12}$$

and $\varphi$ is a vector of regression which consists of measured values of inputs and outputs

$$\varphi^T(t) = \left[-y(t-1), \cdots\cdots, -y(t-n_a), u(t-1), \cdots\cdots, u(t-n_b-1)\right] \tag{13}$$

The model given in (11) presents an accurate description of the system. However, in this expression the vector of system parameters $\theta$ is unknown. It is important to determine it by using available data in signal samples at system output and input. For that purpose a model of the system is supposed

$$\hat{y}(t) = \varphi^T(t)\hat{\theta}(t-1) \tag{14}$$

For the RLS algorithm to be able to update the parameters at each sample time, it is necessary to define an error. The model prediction error, $\varepsilon(t)$ is a key variable in RLS algorithm and is defined as

$$\varepsilon(t) = y(t) - \hat{y}(t) = y(t) - \varphi^T(t)\hat{\theta}(t-1) \tag{15}$$

The error $\varepsilon(t)$ is the difference between the system output and the estimated model output. This model prediction error is used to update the parameter estimates as

$$\hat{\theta}(t) = \hat{\theta}(t-1) + P(t)\varphi(t)\varepsilon(t) \tag{16}$$

where the estimator covariance matrix $P(t)$ is updated using

$$P(t) = \frac{1}{\lambda}P(t-1)\left[I_p - \frac{\varphi(t)\varphi^T(t)P(t-1)}{\lambda + \varphi^T(t)P(t-1)\varphi(t)}\right] \tag{17}$$

where the subscript '$p$' is the dimension of the identity matrix, $p=n_a+n_b+1$, $\lambda$ is the forgetting factor, $0<\lambda\leq 1$. The property of the forgetting factor, $\lambda$, is that $\lambda$ controls the speed of parameter convergence: $\lambda=1$ yields the slowest speed, but provides the best robustness towards noise, and decreasing values of $\lambda$ result in increasing speed of parameter convergence. In general, choosing $0.98<\lambda<0.995$ gives a good balance between convergence speed and noise susceptibility (Alexander, 2001).

Application of RLS method demands supposition of the initial values of $P(t)$ and $\hat{\theta}(t)$. The technique which is chosen as estimates and then allowed to settle to their final values as the program goes through several iterations. There is no unique way to initialize the algorithm. One suggestion is using a supposition that the system is an integrator of the first order with unit gain to set $\hat{\theta}(0)$. While, a standard choice of $P(0)$ is the unit matrix scaled by a positive scalar $\alpha$, *(i.e.* $P(0)=\alpha I_p$ *)*, where $\alpha$ is recommended to be chosen $1<\alpha<10^3$ depending on the existence of prior knowledge about the system parameters (Wellstead, 1991).

A square wave perturbation signal with a frequency of approximately 0.2 of the system bandwidth ensures that most of the square wave power, associated with the first three harmonic components, is inside the system bandwidth. A square wave perturbation signal with a frequency of $f$=0.05Hz that is superimposed on a step signal was applied to the system input. The RLS algorithm is implemented for experimental tests using SIMULINK and real-time windows target.

Table 3: MSE for different estimated models.

| Order | 1st | 2nd | 3rd | 4th | 5th |
|---|---|---|---|---|---|
| MSE | 0.00617 | 0.00368 | 0.00357 | 0.00355 | 0.00392 |

For comparison purposes, Table 3 shows the MSE calculated for different model orders. Third order

model appears to be suitable for describing the system. Further increase in the model order brought no significant improvement.

The performance of the estimated parameters and the model output error for a third-order model are shown in Fig. 5. Estimated parameters converge after a certain time. The speed of parameter convergence depends on the forgetting factor used. Faster parameter convergence can be obtained if the value of the forgetting factor is reduced, but noise amplification. The measured system output and predicted model output is shown in Fig. 6. It can be seen that both output signals are in good agreement.

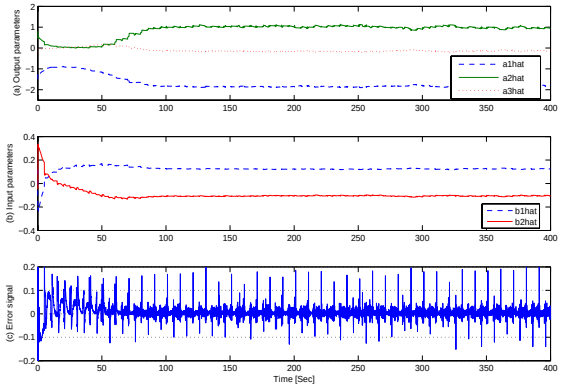



Figure 5: Open-loop estimated parameters for 3rd order model.

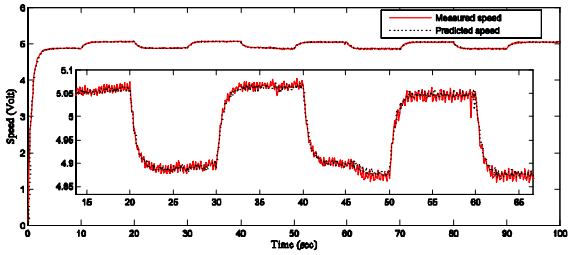



Figure 6: Measured and predicted speed for 3rd order model.

## 4.2 Closed-loop System Identification

Closed-loop identification using direct method is considered in this section. Knowledge of the controller or the nature of the feedback is not a certain requirement. A local-optimal controller (Abou-Zayed, 2008) that provides stable closed-loop servo operation is implemented, using SIMULINK and real-time windows target.

The estimated parameters for the third-order model are shown in Fig. 7. The parameters converge faster than open-loop identification. Further, the variations in the estimated parameters are smaller than that obtained from the open-loop identification. This phenomenon is due to the closed-loop feedback control since the local-optimal controller filters high frequency signal components and limits the bandwidth.

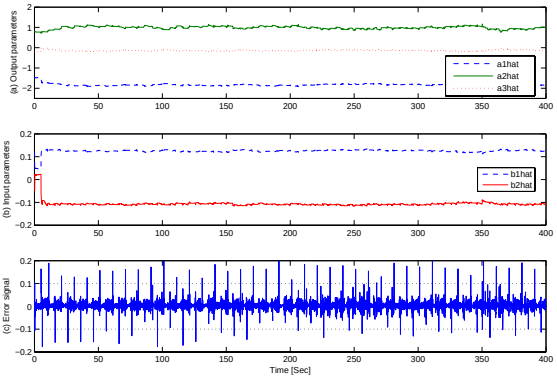



Figure 7: Closed-loop estimated parameters for 3rd order model.

Table 4: Estimated parameters of third-order model for open-loop and closed-loop experiments.

| Parameters | Open-loop | | Closed-loop | |
|---|---|---|---|---|
| | Magnitude | Variation | Magnitude | Variation |
| $a_1$ | -1.821 | 0.204 | -1.855 | 0.082 |
| $a_2$ | 1.065 | 0.172 | 0.942 | 0.136 |
| $a_3$ | -0.122 | 0.144 | -0.131 | 0.080 |
| $b_1$ | 0.132 | 0.025 | 0.125 | 0.008 |
| $b_2$ | -0.105 | 0.029 | -0.113 | 0.009 |

Table 4 shows the third-order parameters estimates for both open-loop and closed-loop experiments. It shows smaller variations for all parameters estimated using closed-loop experiment. That comes due to the closed-loop local-optimal control which filters high frequency signal components and limits the bandwidth

# 5 CONCLUSIONS

This paper presents theoretical and experimental identification of a three-mass electromechanical servo system using different algorithms. The aim of

this research is also to highlight some of the more practical implications of plant identification and to describe the well-established algorithm, recursive least squares, used to perform system identification.

On-line open-loop and closed-loop identification of the studied system is discussed. A real-time implementation of the RLS estimator is presented using SIMULINK and real-time windows target. The application of the RLS method is also demonstrated on a real-time experimental set-up such that it is practical and easy to use. A third-order discrete-time linear model is shown to be flexible enough to fit the observations well. It also became apparent that the order of the suitable linear model was lower than the theoretical one. Closed-loop identification gives faster parameters convergence than open-loop identification. Further, the variations in the estimated parameters are smaller than that obtained from the open-loop identification. This phenomenon is due to the closed-loop local-optimal control which filters high frequency signal components and limits the bandwidth.

## ACKNOWLEDGEMENTS

The authors gratefully acknowledge the support of this work by the EPSRC grant EP/C015185/1.

## REFERENCES

Abou-Zayed, U., Ashry, M., & Breikin, T. (2008). *Implementation of local optimal controller based on model identification of multi-mass electromechanical servo system*. Paper presented at the Proceedings of the 27th IASTED International Conference on Modelling, Identification, and Control.

Alexander, C. W., & Trahan, R. E. (2001). A comparison of traditional and adaptive control strategies for systems with time delay. *ISA Transactions, 40*(4), 353-368.

Ljung, L. (1997). *System identification toolbox : for use with MATLAB*. Natick, Mass.: MathWorks Inc.

Ljung, L. (1999). *System identification: theory for the user*. Upper Saddle River, N.J.; London: Prentice Hall PTR: Prentice-Hall International.

Schroeder, M. (1970). *Synthesis of low-peak factor signals and binary sequences with low autocorrelation. IEEE Trans. Inform. Theory, IT-16, 85-89.*

Soderstrom, T. (1989). *System identification. New York; London: Prentice Hall.*

Wellstead, P. E., & Zarrop, M. B. (1991). *Self-tuning systems: control and signal processing. Chichester: Wiley.*

Ziaei, K., & Sepehri, N. (2000). *Modeling and identification of electrohydraulic servos. Mechatronics, 10(7), 761-772.*

Ziegler, J. G., & Nichols, N. B. (1942). *Optimum settings for automatic controllers. Transactions of the ASME, 64, 759-768.*

# FREQUENCY CONTROL FOR ULTRASONIC PIEZOELECTRIC TRANSDUCERS, BASED ON THE MOVEMENT CURRENT

Constantin Voloşencu
*Automatics and Applied Informatics Department, "Politehnica" University of Timisoara*
*Bd. V. Parvan nr. 2, 300223 Timisoara, Romania*
*constantin.volosencu@aut.upt.ro*

Keywords: Control systems, Piezoelectric transducers, Frequency control, High power ultrasonics.

Abstract: This paper provides a method for frequency control at the ultrasonic high power piezoelectric transducers, using a feedback control systems based on the first derivative of the movement current. This method assures a higher efficiency of the energy conversion and greater frequency stability. A simulation for two kinds of transducer model is made. The method is implanted on a power electronic generator. Some transient characteristics are presented.

## 1 INTRODUCTION

Piezoelectric transducers (Gallego-Juarez, 1989) have proved their huge viability in the high power ultrasonic applications as cleaning, welding, chemical or biological activations and other for many years (Hulst, 1972), (Neppiras, 1972). And these applications continue to be of a large necessity.

The power ultrasonic transducers are fed with power inverters, using transistors working in commutation at high frequency (Bose, 1992). A large scale of electronic equipments, based on analogue or digital technology, is used for control in the practical applications (Marchesoni, 1992.).

A good efficiency of the energy conversion in the power ultrasonic equipments is very important to be assured. Different control methods are used in practice to control the signal frequency in the power inverters (Ramos, et. all., 1985), (Fabianski and Palczynski, 1989).

The high power ultrasonic piezoelectric transducers are analysed with complex structures by using equivalent circuits (Lazaro et. all. 1989), starting from Mason's model, implemented on circuit analysis programs (Morris, 1986).

Many frequency control methods, structures and devices, on this field of interest, are patented.

This paper presents a method to control the frequency of the feeding voltage for the piezoelectric transducer. A power amplifier working in commutation at high frequency generates the feeding voltage. The control system is based on a PI controller, which keep at zero the derivative of the movement current. This control assures the maximum of the mechanic power generated by the transducer.

## 2 RELATED WORK

To perform an effective function of an ultrasonic device for intensification of different technological processes a generator should have a system for an automatic frequency searching and tuning in terms of changes of the oscillation system resonance frequency. The article (Khmelev, et all., 2001) presents a system of phase-locked-loop frequency control of ultrasonic generators with automatic resonance frequency searching in the given band of frequencies.

In (Furuichi and Nose, 1981) a driving circuit for ultrasonic tools which uses a piezoelectric transducer to convert ultrasonic electric signals into ultrasonic mechanical vibrations includes a voltage-controlled oscillator which produces an output signal at a frequency that is proportional to an input voltage, a power amplifier stage having its input coupled to the output of the voltage-controlled oscillator.

In (Hasegawa, 2003) an automatic frequency control (AFC) circuit is based on a frequency offset estimating circuit produces a lock signal if a calculated frequency error becomes smaller than a predetermined value.

The dynamic characteristics of a fixed measuring transducer are defined not only by the parameters of its mechanical system and the ability to convert mechanical into electrical energy but also by the properties of the object on which the transducer is mounted and by the mounting rigidity. In the paper (Senchenkov, 1991) the block diagram is discussed of a system that will detect and evaluate the faults in measuring transducers by comparing the amplitude-frequency characteristic obtained by applying electric pulses to the piezoelement of the transducer with the amplitude-frequency characteristics obtained for the transducer in its natural state and at the moment after it has been mounted on an object.

In (Sullivan, 1983) a power supply is provided for an electromechanical device of the type employing ultrasonic frequency vibratory energy for bonding materials. An automatic frequency control varies the output frequency of the power supply until the ratio of the maximum to minimum amplitudes of a standing wave produced in the mechanical vibratory member falls below a pre-set maximum. The power supply frequency is automatically varied to maintain the standing wave ratio below a pre-set value which is deemed to be an acceptable value for efficient transfer of power.

Figure 1 a): A piezoelectric transducer.

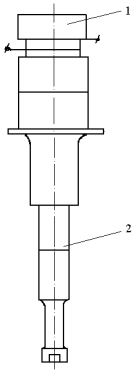



Figure 1 b): Piezoelectric transducer coupled with an energy concentrator.

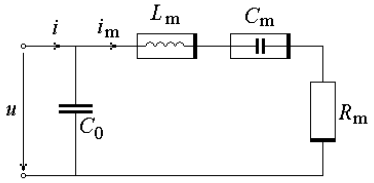



Figure 2: The equivalent circuit of the piezoelectric transducer.

# 3 CONTROL PRINCIPLE

## 3.1 The Equivalent Circuit of the Piezoelectric Transducer

The electrical energy of ultrasonic frequency in the bandwidth of 20-80 kHz is converted in mechanical energy using piezoelectric transducers. A kind of piezoelectric transducer of 100 W is presented in figure 1,a. In practice the transducer (1) is mechanical coupled with an energy concentrator (2), as is presented in figure 1,b).

The ultrasonic piezoelectric transducers have the equivalent electric circuit from figure 2. In this circuit there is emphasized the mechanical part, seen as a series RLC circuit, with the equivalent parameters $R_m$, $L_m$ and $C_m$, which are nonlinear, depending on the transducer load.

The current through the mechanical part $i_m$ is the movement current. The input capacitor $C_0$ of the transducer is consider as a constant parameter.

The equations (1) are describing the time variation of the signals and the mechanical parameters.

$$
\begin{aligned}
u_{Lm} &= \frac{d\varphi_{Lm}}{dt} = \frac{dL_m}{dt} i_m + L_m \frac{di_m}{dt} \\
i_{Cm} &= \frac{dq_{Cm}}{dt} == \frac{dC_m}{dt} u_{Cm} + C_m \frac{du_{Cm}}{dt} \\
i_{Cm} &= \frac{dq_{Cm}}{dt} == \frac{dC_m}{dt} u_{Cm} + C_m \frac{du_{Cm}}{dt} \\
R_m &= \frac{du_{Rm}}{di_m} \\
i &= i_0 + i_m \\
u &= u_{Lm} + u_{Cm} + u_{Rm} \\
C_0 \frac{du}{dt} &= i_{C0}
\end{aligned}
\qquad (1)
$$

where φ is the magnetic flux through the mechanical inductance $L_m$ and $q$ is the electric load over the mechanical capacitor $C_m$.

The piezoelectric traducer has a frequency characteristic of its impedance $Z$ with a series and a parallel resonance, as it is presented in figure 3.

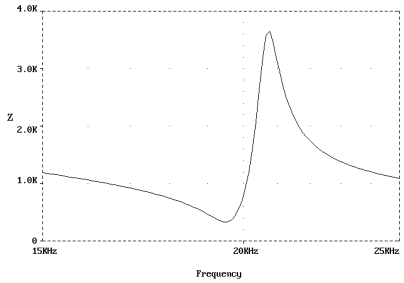



Figure 3: The frequency characteristic of the transducer impedance.

The movement current $i_m$ has the frequency characteristics from figure 4.

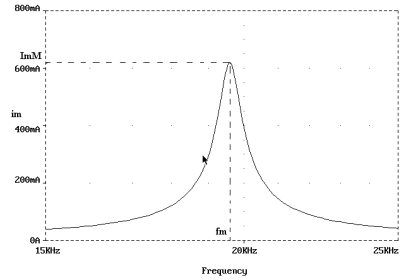



Figure 4: The frequency characteristic of the transducer impedance.

The maximum mechanical power developed by the transducer is obtained when it is fed at the frequency $f_m$, were the maximum movement current $i_m=I_{mM}$ is obtained. Of course, the maximum of the movement current $i_m$ is obtained when the derivative of the absolute value of the movement current *dim* is zero:

$$\dim(t) = \frac{d\,|\,i_m\,|}{dt} = 0 \tag{2}$$

So, a frequency control system, functioning after the error of the derivative movement current may be developed, is using a PI frequency controller, to assure a zero value for this error in the permanent regime.

## 3.2 The Frequency Control System

The block diagram of the frequency control system based on the above assumption is presented in figure 5.

A power amplifier AP, working in commutation, at high frequencies, feeds a piezoelectric transducer E, with a rectangular high voltage $u$, with the frequency $f$. An output transformer T assures the high voltage $u$ for the ultrasonic transducer E. A command circuit CC assures the needed command signals for the power amplifier AP. The command signal $u_c$ is a rectangular signal, generated by a voltage controlled frequency generator GF_CT. The rectangular command signal $u_c$ has the frequency $f$ and equal durations of the pulses. The frequency of the signal $u_c$ is controlled with the voltage $u_f$*. The signal $u_f$* to control the frequency $f$ of the transducer is provided by the frequency controller RG-f.

The frequency control system from figure 4 is based on the derivative movement current error $e_{dim}$:

$$e_{dim} = \dim^* - \dim \tag{3}$$

as the difference between the reference value *dim**=0 and the computed value of the derivative *dim*.

A PI controller is used to control the frequency, with the following transfer function:

$$u_f^*(s) = K_R\left(1 + \frac{1}{T_R s}\right)e_{dim}(s) \tag{4}$$

The frequency controller is working after the error of the derivative of the movement current $e_{dim}$.

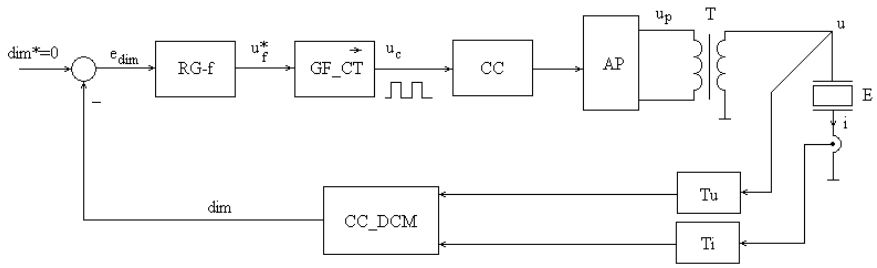



Figure 5: The block diagram of the frequency control system.

The derivative of the movement current *dim* is computed using a circuit CC_DCM, where $C_0$ is the known constant value of the capacitor from transducer input and *u* and *i* are the measured values of the transducer voltage and current.

The voltage upon the transducer *u* and the current *i* through the transducer are measured using a voltage sensor Tu and respectively a current sensor Ti.

# 4 MODELING AND SIMULATION

Some models for different parts of the block diagram from figure 5 were developed to test the control principle by simulation.

Two models are tested for the transducer. In the first model the parameters of the mechanical part are considered with a static value and a dynamical variation. In the second model the electromechanical transducer is considered coupled with a mechanical concentrator.

Approximating the relations (1), the following relations are used to model the behaviour of the piezoelectric transducer:

$$\begin{aligned} u_{Lm}(s) &= sL_m(s)i_m(s) + sL_m(s)i_m(s) \\ i_{Cm}(s) &= sC_m(s)u_{Cm}(s) + sC_m(s)su_{Cm}(s) \\ u_{Rm}(s) &= \frac{1}{s}R_m[si_m(s)] \\ u(s) &= u_{Lm}(s) + u_{Cm}(s) + u_{Rm}(s) \end{aligned} \tag{5}$$

The movement current $i_m$ (*s*) is modelled, based on the above relations, with the following relation:

$$i_m(s) = \frac{1}{s}\left[\frac{1}{L_m} \times \left(u(s) - \frac{1}{s}.\frac{1}{C_m} \times i_m(s) - R_m \times i_m(s)\right)\right] \tag{6}$$

The block diagram of the movement current model is presented in figure 6.

The mechanic parameters from the above relations have the variations given by relations (7), in the vicinity of the stationary points $R_{m0}$, $L_{m0}$ and $C_{m0}$.

$$\begin{aligned} R_m &= R_{m0} + \Delta R_m \\ L_m &= L_{m0} + \Delta L_m \\ C_m &= C_{m0} + \Delta C_m \end{aligned} \tag{7}$$

A second model is taken in consideration. The transducer is considered coupled with the concentrator and the equivalent circuit is presented in figure 7.

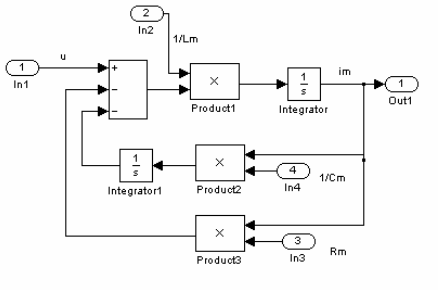



Figure 6: The block diagram, model for the mechanical part of the piezoelectric transducer.

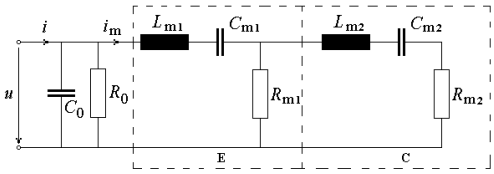



Figure 7: Equivalent circuit of the transducer with concentrator.

In this model there is a series RLC with the parameters $L_{m1}$, $C_{m1}$ and $R_{m1}$ for the transducer T and a series RLC circuit with the parameters $L_{m2}$, $C_{m2}$ and $R_{m2}$ for the concentrator C, coupled in cascade.

The parts of the control block diagram are modelled using Simulink blocks. A transient characteristic of the frequency error $e_{dim}$ from the control system is presented in figure 8.

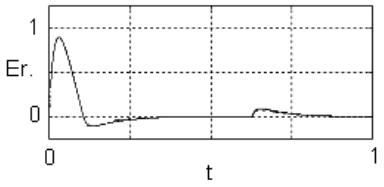



Figure 8: Transient characteristic for the frequency control system, obtained by simulation.

The simulation is made considering for the first model the variation with 10 % at the transducer parameters. The deviation in frequency is eliminated fast. The frequency response has a small overshoot.

# 5 IMPLEMENTATION AND TEST RESULTS

The frequency control system is developed to be implemented using analog, high and low power, circuits, for general usage. The power amplifier AP is realized using four power MOSFET transistors, in a full bridge, working in commutation at high frequency. The voltage controlled frequency generator GF_CT is realized using a phase lock loop PLL circuit and a comparator. The computing circuit CC_DCM, which implements the relations and the frequency controller RG-f are realized using analogue operational amplifiers. The transformer T is realized using ferrite cores.
The electronic generator is presented in figure 9.

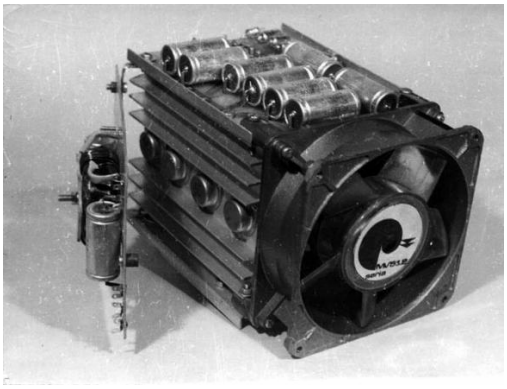


Figure 9: The electronic generator.

In the following figures some transient signal variations of the control system are presented.

The pulse train of the command voltage $u_c$ is presented in figure 10.

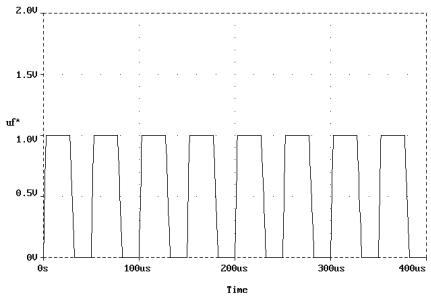



Figure 10: Examples of sensor impulse trains.

The output voltage of the power amplifier is presented in figure 11.

The voltage $u$ over the piezoelectric transducer is presented in figure 12.

The measured movement current $i_m$ is presented in figure 13.

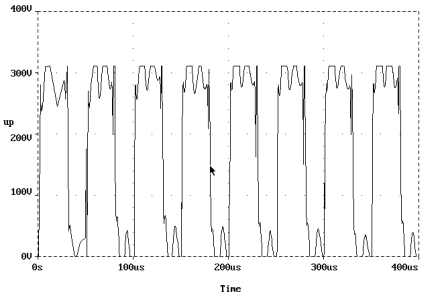



Figure 11: The output voltage.

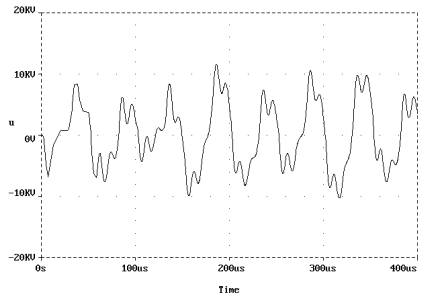



Figure 12: The transducer voltage.

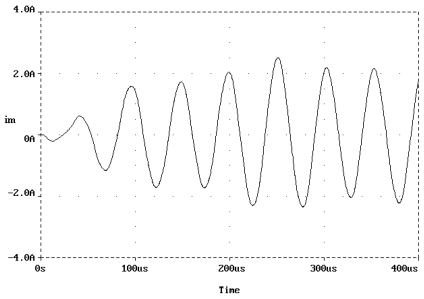



Figure 13: The movement current.

# 6 CONCLUSIONS

In this paper a method to control the frequency of the piezoelectric ultrasonic transducers based on the movement current through the mechanical part of the equivalent circuit of the transducer is presented. The principle of this method is to assure the maximum mechanical power developed by the transducer, based on the measured transducer's voltage and current, controlling the feeding voltage frequency, as the derivative of the movement current to be zero.

The frequency control system was modelled and simulated using Matlab and Simulink. Two models for the mechanical part of the transducer are chosen. Two different regimes for the time variations of the mechanical parameters of the transducer was chosen and tested. A Simulink model and a simulation result are presented. The simulation results have proven that the control principle developed in this paper gives good quality criteria for the output frequency control.

The control system is implemented using a power inverter with transistors working in commutation at high frequencies and analogue circuits for command. Transient characteristics of the control systems are presented.

The frequency control system may be developed for piezoelectric transducers in a large scale of constructive types, powers and frequencies, using general usage analogue components, at a low price, with good control criteria.

## REFERENCES


Bose, B.K., 1992. Evaluation of Modern Power Semiconductor Devices and Future Trends of Converters, In *IEEE Trans. on Industry Applications*, march/april, vol.28, no. 2.

Gallego-Juarez, J.A., 1989. Piezoelectric Ceramics and Ultrasonic Transducers, In *J. Phys. Sci. Instrum.* (U.K.), oct., vol. 22, no. 10.

Hulst, A.P., 1972. Macrosonics in industry 2. Ultrasonic welding of metals, In *Ultrasonics*, Nov.

Fabianski, P., Palczynski, L., 1989. Power Inverter with Self-Tuning Output Frequency for Ultrasonic Cleaning System, In EPE'89, 3-rd European Conference on Power Electronics and Applications, Aachen, Germany, oct.

Khmelev, V.N., Barsukov, R.V., Barsukov, V., Slivin, A.N., Tchyganok, S.N., 2001. System of phase-locked-loop frequency control of ultrasonic generators, In *Electron Devices and Materials, 2001. Proceedings. 2nd Annual Siberian Russian Student Workshop on***.**

Lazaro, O.J.C., San Sanche, P.T., Gallego-Juarez, J.A., 1989. Analysis of an ultrasonic transducer with complex structure by using equivalent circuits, In Ultrasonics International, Conference Proceedings, Madrid, Spain.

Marchesoni, M., 1992. High-Performance Current Control Techniques for Applications to Multilevel High-Power Voltage Source Inverters, In *IEEE Trans. on Power Electronics*, Jan.

Morris, A.S., 1986. Implementation of Mason's model on circuit analysis programs, In *IEEE Transactions on ultrasonics, ferroelectric and frequency control*, vol. UFFC-33, no. 3.

Mori, E., 1989. High Power Ultrasonic Wave Transmission System, In *J. Inst. Electron. Inf. Commun. Eng.*, vol. 72, no. 4, April.

Neppiras, E.A., 1972. Macrosonics in industry, 1. Introduction. In *Ultrasonics,* Jan.

Ramos, F.A., Montoya, V.F., Gallego-Juarez, J.A., 1985. Automatic system for dynamic control of resonance in high power and high Q ultrasonic transducers, In *Ultrasonics*, July.

I. K.Senchenkov, I.K., 1991. Resonance vibrations of an electromechanical rod system with automatic frequency control, In *International Applied Mechanics*, Vol. 27, No. 9/ Sept., Springer, N. Y.

Furuichi, S., Nose, T., 1981. Driving system for an ultrasonic piezoelectric transducer, *U.S. patent 4271371*.

Hasegawa, O., 2003. Automatic frequency control circuit, *U. S. Patent 6571088*.

Sullivan, R.A., 1983. Power supply having automatic frequency control for ultrasonic bonding, *U. S. Patent 4389601*.

# A UNIFYING POINT OF VIEW IN THE PROBLEM OF PIO
## *Pilot In-the-loop Oscillations*

Vladimir Răsvan, Daniela Danciu and Dan Popescu
*Department of Automatic Control, University of Craiova*
*13 A. I. Cuza Str., RO-200585 Craiova, Romania*
{*vrasvan, daniela, dpopescu*}*@automation.ucv.ro*

Keywords: Oscillations, Feedback structure, Robustness.

Abstract: The paper starts from the problem of PIO (Pilot-In-the-loop Oscillations), a major problem in aircraft handling and control, where the idea of the *feedback as hidden technology* is basic. The real phenomenon called PIO is modeled by a feedback structure where the pilot acts as one of the components of the loop and has to be modeled accordingly. PIO are in fact self-sustained oscillations and usually are divided into three convenient categories that are based on the nature of the pilot and vehicle dynamics behavior models and analysis needed for their explanation. Category I PIO are essentially linear while Category II PIO are quasi-linear and typically associated with rate limiting. Category III PIO are fully nonlinear and non-stationary. Since PIO II are mostly tackled *via* various robustness approaches starting from linear models, the paper strives for a unifying approach which is illustrated accordingly.

## 1 BASICS AND PROBLEM STATEMENT

According to the standard terminology of the field, PIO (Pilot Induced Oscillations, Pilot In-the-loop Oscillations, Pilot Involved Oscillations) are sustained or uncontrollable oscillations resulting from the effort of the pilot to control the aircraft, hence they can be considered as a closed loop destabilization of the aircraft-pilot loop (Anon., 2000),(McRuer et al., 1996). Even from this remarkably short definition it appears that the real phenomenon called PIO can be modeled by a feedback structure where the pilot acts as one of the components of the loop and has to be modeled accordingly. As mentioned in (McRuer et al., 1996) the study of the aeronautical history reveals a remarkably diverse set of severe PIOs as exemplified by the listings of "famous PIOs" (McRuer, 1994),(Klyde et al., 1995).

The feedback control character of PIOs was recognized almost from the outset because the aircraft left alone did not exhibit such oscillations. Once recognized as oscillations within a feedback system context, mathematical models were developed and used to describe the pilot's dynamic actions as a controller and active participant in PIOs.

**A.** Detailed analytical studies of past PIO incidents (see e.g. the references from (McRuer et al., 1996)) relied on pilot behavioral models and closed loop analysis procedures to understand and rationalize phenomena. Moreover in some cases pilot vehicle behavioral models were applied to design and assess changes to the effective vehicle to alleviate the PIO potential. Based on these results it is useful to divide PIOs into categories that reflect the analytical and pilot modeling tools. There were identified three categories of PIO as follows:

- Category I - Essentially Linear Pilot Vehicle System Oscillations: the element characteristics are essentially linear and the pilot behavior is linear (except, possibly, for simple gain shaping in series with the pilot).
- Category II - Quasi-Linear Pilot Vehicle Systems with Series Rate or Position Limiting. Rate limiting, either as a series element or as a rate limited surface actuator modifies the Category I situation by adding what is called (non-rigorously) an amplitude dependent lag and by setting the limit cycle magnitude.
- Category III - Essentially Non-Linear Pilot Vehicle System Oscillations with transitions: they fundamentally depend on nonlinear transitions in either the effective control element or in the pilot behavioral dynamics.

**B.** Most of the available information shows that mainly PIO I and PIO II were considered and analyzed due to the complexity of PIO III which never-

theless have been recognized as quite rare and arising from PIO I and PIO II; consequently a consensus has been established in the PIO community that PIO III proneness may be blocked by blocking PIO I and PIO II proneness. With respect to this several PIO I and PIO II criteria have been elaborated in USA and in Europe. These criteria are viewed as sufficient conditions ensuring that the feedback system pilot-aircraft is PIO free. Most of them are obtained in the linear case i.e. for PIO I. Here several remarks are necessary. It is a trivial fact that for linear systems there exist necessary and sufficient conditions for stability what means also absence of self sustained oscillations. The PIO criteria are only sufficient conditions but they are conceived as to ensure some kind of robustness with respect to system's uncertainties. Indeed the presence of uncertainties is quite obvious. There are first the uncertainties of aircraft modeling - aerodynamic forces and coefficients depending of the flight envelope parameters - but also those of pilot modeling which depend on several modeling assumptions.

On the other hand, as already mentioned, PIO II are associated to quasi-linear models where rate and position limiters are active. The limiters are modeled as structures containing saturation nonlinear functions; the quite recent models which are based on Integral Quadratic Constraints (Megretski and Rantzer, 1997),(Megretski, 1997) take into account the simplest remark that the saturation nonlinearity (fig. 1) may be "embedded" in the larger class of the sector restricted nonlinearities (fig. 2)

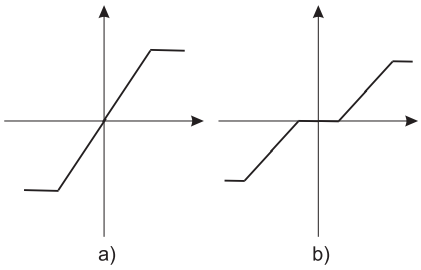



Figure 1: Saturation nonlinearities.

As known, the properties of the sector restricted nonlinearities may be expressed under the form of some quadratic constraints - see (Megretski and Rantzer, 1997) but also the pioneering paper (Yakubovich, 1967) as well as the monograph (Răsvan, 1975). This embedding of the nonlinear function in a larger class speaks about allowing some uncertainty concerning the nonlinearity; additionally, if the stability results are valid uniformly for the entire class of non linearities, some robustness is ensured.

Consequently, it is not by chance that an important class of methods associated to PIO II originate from robustness approaches. Moreover, one can find there such standard methods of the absolute stability as the Liapunov function of the form "quadratic form of the state variables plus integral of the nonlinear function" or the Popov frequency domain inequality (Anon., 2000).

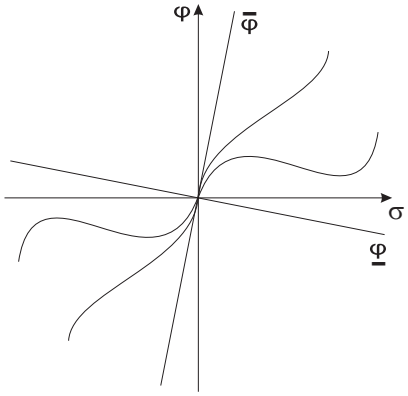



Figure 2: Sector restricted nonlinearity.

The above considerations show that it is not without interest *to discuss PIO I and PIO II within a unified context of robustness* in the sense that the robustness restrictions introduced in the totally linear case (PIO I) should be taken into account in the quasi-linear (PIO II) case. As an at-hand example, the frequency domain restrictions of the Neal-Smith criterion (Neal and Smith, 1971) should be reflected in a Popov like frequency domain inequality.

The present paper will demonstrate and motivate the above sketched approach and what remains is organized as follows. Firstly the basic feedback structure is presented in the context of fully linear models accounting for PIO I. It is then shown how rate limiters occur in the loop - the PIO II onset - and the new structure of a feedback nonlinear system with a sector restricted nonlinearity. The linear subsystem is then identified and a frequency domain inequality is then formulated. This inequality has to be valid for the frequency domain characteristic as resulting from a PIO I criterion; if this holds we may say that PIO I gives "some insurance" for PIO II. Next a specific case will be discussed to illustrate the principle and conclusions together with suggestions for future research and tests are given.

## 2 ROBUSTNESS VERSUS ABSOLUTE STABILITY

The analysis of the models of (Anon., 2000),(McRuer et al., 1996),(Klyde et al., 1996),(Klyde and Mitchell,

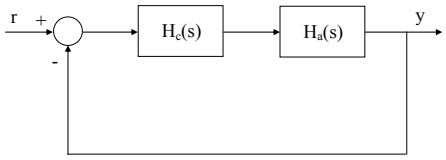



Figure 3: Basic linear feedback structure.

2005) which deal with longitudinal PIO will suggest the general feedback structure of fig. 3.

We denoted there by $H_a(s)$ the transfer function of the "uncontrolled plant" which might be some longitudinal or lateral motion of the aircraft; by $H_c(s)$ we denoted the "controller" which in this man/machine system might be some pilot model (for instance the so-called *synchronous pilot* is just a gain - see (Klyde et al., 1996)).

Some remarks are necessary from this very beginning. Since various assumptions on pilot behavior may require *pole/zero cancelation*, only LHP (left hand plane) i.e. stable poles and zeros may be canceled, otherwise uncontrollable unstable modes will appear. This is particularly true for the so-called *crossover model* where we have

$$H_c(s)H_a(s) \equiv \frac{K}{s}\mathrm{e}^{-\tau s} \tag{1}$$

and this clearly implies pole/zero cancelation. Since it is well known that modern fighters become unstable for high speed points of the flight envelope, they are made stable by additional stabilizing feedback - the SAS (Stability Augmentation System). Equality (1) avoids unstable pole/zero cancelation only if the SAS is active[1].

If the limiters are to be considered, the system of fig. 3 will become a standard feedback control structure with a nonlinear (saturated) actuator (fig. 4)

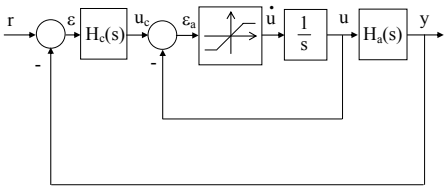



Figure 4: Feedback structure with rate limiter.

In order to obtain the standard structure of the absolute stability problem, we consider the state realizations of the two proper rational transfer functions $H_a(s)$ and $H_c(s)$ embedded in the structure of fig. 4

[1]This explains also in some way the X-15 landing flare PIO evoked in (Klyde et al., 1996) since it is mentioned there that the "pitch damper was off", the pitch damper being the SAS of the channel

$$\begin{aligned}
&\dot{x}_a = Ax_a + bu \;,\; y = c^T x_a + h_0 u \\
&\dot{u} = \varphi(\varepsilon_a) \;,\; \varepsilon_a = u_c - u \\
&\dot{x}_c = A_c x_c + b_c \varepsilon \;,\; \varepsilon = r(t) - y \\
&u_c = f_c^T x_c + h_c \varepsilon
\end{aligned} \tag{2}$$

which becomes

$$\begin{aligned}
&\dot{x}_a = Ax_a + bu \\
&\dot{x}_c = -b_c c^T x_a + A_c x_c - h_0 b_c u + b_c r(t) \\
&\dot{u} = \varphi(\sigma) \\
&\sigma = -h_c c^T x_a + f_c^T x_c - (1 + h_c h_0)u + h_c r(t)
\end{aligned} \tag{3}$$

For $r(t) \equiv 0$ what means the system is considered in deviations with respect to some steady state (equilibrium) the feedback structure of fig. 5 is obtained

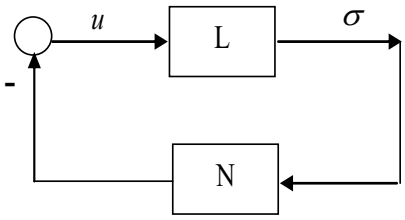



Figure 5: Absolute stability feedback structure.

The linear subsystem is described by the controlled system of ordinary differential equations with linear output

$$\begin{aligned}
&\dot{x}_a = Ax_a + bu \\
&\dot{x}_c = -b_c c^T x_a + A_c x_c - h_0 b_c u \\
&\dot{u} = -\mu(t) \\
&\sigma = -h_c c^T x_a + f_c^T x_c - (1 + h_c h_0)u
\end{aligned} \tag{4}$$

in feedback connection with the nonlinear static block

$$\mu = -\varphi(\sigma) \tag{5}$$

The transfer function of (4) is

$$H(s) = \frac{\tilde{\sigma}(s)}{\tilde{\mu}(s)} = \frac{1}{s} + \frac{1}{s} H_c(s)H_a(s) \tag{6}$$

and the characteristic equation clearly has a zero root. This might be the simplest critical case of the absolute stability, but if (1) holds the case corresponds to the non-simple zero root - the most special critical case, that was studied separately of the other ones, due to its specific problems; moreover the presence of the delay in (1) will complicate the approach (Răsvan, 1975); we give here an adaptation of Theorems 6.1 and 7.2 of (Răsvan, 1975)

**Theorem 1.** *Consider the system of fig. 5 where the linear subsystem has its transfer function of the form (6) with $H_c(s)H_a(s)$ being a meromorphic function - ratio of quasi-polynomials; the denominator has a simple zero root and all other roots with negative real parts. The nonlinear function $\varphi$ is subject to the following conditions*

$$\begin{aligned} &\sigma\varphi(\sigma) > 0 \ (\sigma \neq 0) \ , \ \varphi(0) = 0 \ , \\ &\lim_{\sigma\to\pm\infty} \int_0^\sigma \varphi(\lambda)\mathrm{d}\lambda = \infty \end{aligned} \tag{7}$$

*Assume that*

$$1 - \lim_{s\to 0_+} H_c(s)H_a(s) > 0 \tag{8}$$

*and also that the frequency domain inequality*

$$1 + \mathrm{Re}\, H_c(\imath\omega)H_a(\imath\omega) > 0 \tag{9}$$

*holds for all $\omega > 0$. Then the system has the absolute stability property.*

# 3 A SIMPLE APPLICATION: ROBUSTNESS OF THE NEAL - SMITH CRITERION

We shall consider here one of the cases of (Klyde and Mitchell, 2005), the so-called crossover PVS (Pilot Vehicle System) model with $H_c(s)H_a(s)$ as in (1), the parameters being chosen to satisfy

- crossover frequency $\omega_c = 1.4$ rad/sec,
- neutral stability frequency $\omega_0 = 1.73$ rad/sec,
- phase margin $\Phi_c = 20^o$,
- gain margin $M_c = 4.45$ dB,
- peak magnification ratio 3.39 at 1.48 rad/sec

Remark that all these performance indicators are stated in the frequency domain and three of them deal with open loop characteristics ($\omega_c$,$\Phi_c$,$M_c$) while the other are concerned with the closed loop properties. Worth mentioning that there are only two free parameters of the model (1) while five conditions are imposed. We may check them as follows. Since $\omega_c$ corresponds to 0 dB in the *gain/log* characteristic, it follows that $K/\omega_c = 1$, hence $K = \omega_c$. On the other hand, since the phase margin is $20^o$, the phase should be $-160^o$ at the crossover frequency what will give $\tau\omega_c = 70^o$ hence $\tau = 7\pi/(18\omega_c)$. With this (1) is completely determined and we have to check the other properties. In order to verify the gain margin we need the frequency of phase reversal (when the phase equals $-180^o$. It follows easily that $\tau\omega_\pi = \pi/2$, therefore

$$\omega_\pi = \frac{9}{7}\omega_c \ , \ A(\omega_\pi) = \frac{K}{\omega_\pi} = \frac{\omega_c}{\omega_\pi} = \frac{7}{9} \tag{10}$$

hence $M_c = 20\lg(9/7) = 2.18$ dB. The condition on the neutral stability frequency $\omega_0$ has to indicate a tolerable increase of the gain provided the time lag is kept constant or, conversely, a tolerable increase of the time lag provided the gain is kept constant. The closed loop characteristic equation is

$$s + K\mathrm{e}^{-\tau s} = 0 \tag{11}$$

and the neutral stability will require a pair of zeros of (11) on $\imath\mathbb{R}$ hence the conditions

$$K\cos\ \omega_0\tau = 0 \ , \ \omega_0 - K\sin\ \omega_0\tau = 0 \tag{12}$$

The first equality gives $\omega_0\tau$ which is substituted in the second to obtain the admissible value of $K$ since $\tau$ follows by fixing $\omega_0$. We may continue in this way by checking the other conditions. Our aim however is to check the usefulness of the proposed approach by applying Theorem 1. Considering the transfer function of (1) we check the frequency domain inequality (9)

$$1 + \mathrm{Re}\,\frac{K\mathrm{e}^{-\imath\omega\tau}}{\imath\omega} = 1 - K\tau\frac{\sin\ \omega\tau}{\omega\tau} > 0 \ , \ \forall\omega$$

and for it fulfilment it is necessary and sufficient to have $1 - K\tau > 0$ which is exactly (8). Taking into account the computations of the linear case we find that $1 - K\tau$=$1 - \omega_c\tau$=$1 - 7\pi/18 < 0$!. This is quite unpleasant and it deserves some comment. Condition (8) accounts for the so-called *limit stability property* (Răsvan, 1975) - a necessary condition for absolute stability within the sector $(0, \overline{\varphi})$ - exponential stability for linear characteristics within arbitrarily small sector $(0, \varepsilon)$. A more general and less restrictive necessary condition might be the so called *minimal stability* (Popov, 1973) which requires stability for a single linear characteristic within the sector; nevertheless, if limit stability fails this will require a linear characteristic within a sector $(\underline{\varphi}, \overline{\varphi})$ with $\underline{\varphi} > 0$ and this is unacceptable since the saturation nonlinearity belongs to the sector $(0, \overline{\varphi})$.

Coming back to the condition $1 - \omega_c\tau > 0$ which does not hold, it follows that robustness assumed in the linear case is not enough to ensure it in the PIO II (system with rate limiter) case. If we require from the beginning $\omega_c\tau < 1$, the phase at the crossover frequency will be larger than $-(1 + \pi/2)$ rad hence the phase margin has to be larger than $\pi/2 - 1$ rad i.e. $\approx 33^o$ - a result that was at some extent expected.

# 4 CONCLUSIONS

This paper is demonstrating a point of view that seemed very natural when absolute stability i.e. robust global asymptotic stability for systems with sector restricted nonlinear functions was investigated. This point of view is that robust stability of linear systems should imply the same property for nonlinear systems also, at least for those with sector restricted nonlinearities. Such a point of view is transparent throughout all research concerning the so called Aizerman and Kalman problems (Popov, 1973),(Răsvan, 1975) and geometric similarities of the Nyquist and Popov frequency domain criteria strengthened it. Stating it, obviously is not enough; this *position paper* is pointing to critical analysis and further research, mainly application oriented. We have chosen the field of aircraft oscillations to illustrate this point of view for its practical importance (proved by the intense research activities around PIO problem) as well as for its feedback-based modeling of the dynamics: control appears here as a genuine *hidden technology* and hidden paradigm.

Since the field of aircraft dynamics and handling qualities has very strict requirements and procedures, the amount of the necessary research appears to be high and with a certain degree of complexity. The point of view stated here is to be applied possibly to all cases of PIO I i.e. corresponding to fully linearized systems; for the entire set of criteria, see (Anon., 2000),(Klyde et al., 1995). But for each criterion one may wish to consider several cases of pilot models. For *all these cases* we have to consider the PIO II i.e. the nonlinear, rate limited counterpart. But, besides the comparison of the PIO I criteria and of their nonlinear counterpart, a comparison with the other PIO II criteria, obtained independently of the approach presented in this paper is also necessary.

All this analysis and various comparison of the criteria contain the necessary amount of critical assessment of the present position paper proposal. To this we add the specific PIO approach in aircraft studies: conversion in a checkable form and application on “real data” stored in the aviation databases. Nevertheless it is hoped to follow the approach described here in the next research on other PIO criteria.

# REFERENCES


Anon. (2000). *Flight Control Design - Best Practices*. NATO-RTO Technical Report 29, December 2000.

Klyde, D. H., McRuer, D. T., and Myers, T. T. (1995). Unified pio theory vol.i: Pio analysis with linear and nonlinear effective vehicle characteristics, including rate limiting. Technical Report WL-TR-96-3028, AIAA.

Klyde, D. H., McRuer, D. T., and Myers, T. T. (1996). Pio analysis with actuator rate limiting. Paper 96-3432-CP, AIAA.

Klyde, D. H. and Mitchell, D. G. (2005). A pio case study - lessons learned through analysis. Paper 05-661-CP, AIAA.

McRuer, D. T. (1994). Pilot induced oscillations and human dynamic behavior. Technical report CR-4683 December 1994, NASA.

McRuer, D. T., Klyde, D. H., and Myers, T. T. (1996). Development of a comprehensive pio theory. Paper 96-3433-CP, AIAA.

Megretski, A. (1997). Integral quadratic constraints for systems with rate limiters. Technical Report LIDS-P-2407, Massachussets Institute of Technology, Cambridge MA.

Megretski, A. and Rantzer, A. (1997). System analysis via integral quadratic constraints. *IEEE Transactions on Automatic Control*, 42(6):819–830.

Neal, T. P. and Smith, R. E. (1971). A flying qualities criterion for the design of fighter flight control systems. *Journal of Aircraft*, 8(10):803–809.

Popov, V. M. (1973). *Hyperstability of Control Systems*. Springer Verlag, Berlin-Heidelberg-New York, 1st edition.

Răsvan, V. (1975). *Absolute stability of time lag control systems (in Romanian)*. Editura Academiei, Bucharest, 1st edition.

Yakubovich, V. A. (1967). Frequency domain conditions for absolute stability of control systems with several nonlinear and linear non-stationary blocks (in russian). *Avtomatika i Telemekhanika*, 28(6):5–30.

# ANALYSIS OF REMS GTS ERRORS DUE TO MSL ROVER AND MARTIAN ENVIRONMENT

Eduardo Sebastián, Carlos Armiens and Javier Gomez-Elvira
*Lab. de Robótica y Exploración Planetaria, Centro de Astrobiología, Ctra. Ajalvir Km.4, Torrejón de Ardoz, Spain*
*sebastianme@inta.es*

Keywords: Environmental monitoring, infrared temperature detection and sensor error sources.

Abstract: This paper analyses the external sources of error of the REMS GTS, a contactless instrument to measure ground temperature that is part of the payload of the NASA MSL mission to Mars. Some environment properties such us atmosphere opacity, solar radiance, ground emissivity and rover IR emissions are studied, determining GTS characteristics. The article also proposes a simplified geometrical and thermal model of the rover and environment in order to evaluate and quantify their influence in ground temperature measurements. Finally, the article summarizes simulation results and provides solutions in order to improve sensor accuracy.

## 1 INTRODUCTION

The GTS (Ground Temperature Sensor), one of the REMS (Rover Environmental Monitoring Station) instruments, is mainly dedicated to measure the brightness temperature of Martian surface, using three thermopiles detectors in three infrared IR bands, and looking directly at the ground. The selected channels are 8-14μm, 16-20μm and 14.5-15.5μm.

In general it can be said that two error sources are associated with contactless temperature measurements. On the one hand internal sources, all those related with the sensor, the amplifying electronics, and also errors associated with calibration and sensor degradation. On the other hand we have the errors due to the environment. In the case of the REMS GTS the environment shows difficulties because of the uncertainty in Martian surface emissivity, reflections from the rover and the Sun, and atmosphere absorbance.

The main objective of this paper is to justify the GTS design based on environment restrictions, as well as to obtain a thermal model of the rover and the environment in order to analyse and correct the errors in ground temperature determination.

The paper is organized as follows; section 2 introduces briefly the REMS GTS; section 3 describes the environmental sources of error, including a simplified radiation model of MSL rover. Section 4 shows simulations to evaluate environment influence, using the proposed model. Finally, section 5 summarizes the results.

## 2 REMS GTS DESCRIPTION

REMS is an environmental station designed by the Centro de Astrobiología with the collaboration of national and international partners (CRISA/EADS, Universidad Politécnica de Cataluña (UPC) and Finish Meteorological Institute (FMI)), which is part of the payload of the MSL (Mars Science Laboratory) NASA mission to Mars, figure 1. This mission is expected to launch in the final months of 2009, and mainly consists of a rover with a complete set of scientific instruments.

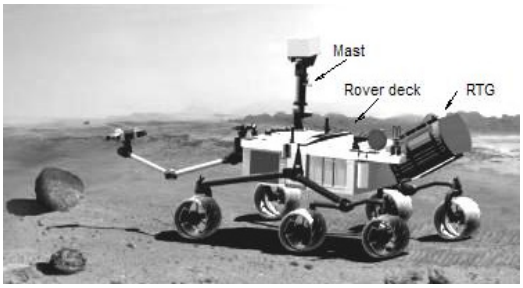



Figure 1: NASA MSL rover.

The rover main body hosts the electronics associated with the whole set of instruments, rover communications and control systems. Additionally, it includes the RTG (Radioactive Temperature Generator) which is the rover energy source. The RTG extra heat is used by rover thermal designers to warm the rover body in order to keep alive the electronics inside.

The GTS shall be mounted in one of the REMS booms, which is placed in the rover mast at 1.6m height and hosts the electronics dedicated to amplify the thermopiles signals. The GTS includes an in-flight calibration system without moving parts, whose main goal is to compensate the sensor degradation due to the deposition of dust over its window (Sebastián and Gomez-Elvira, 2007). To avoid local effects, the GTS focuses a large surface area of around $100^2$m, shown in figure 2, measuring the average temperature. This are is far enough from the rover as to minimize its influence.

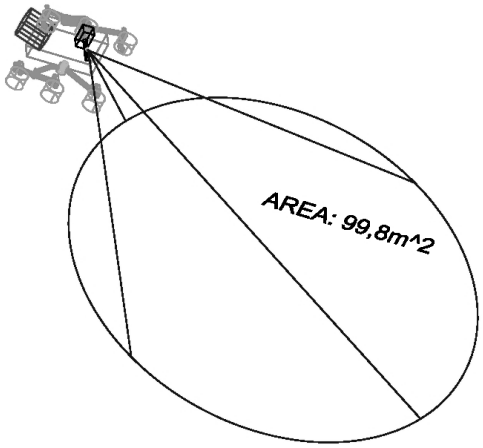



Figure 2: REMS GTS FOV and MSL rover simplified draft.

# 3 GTS ERRORS DUE TO MARTIAN ENVIRONMENT

Contactless temperature measurements are based on the integration of the IR radiation coming from a body. This radiation depends mainly on three factors: The temperature of the focused area, the emissivity $\varepsilon$ of its surface, or what is the same the capacity of the body to emit IR energy, and finally the reflectivity $r$ of its surface, that shows how the body reflects energy coming from the environment.

For the characteristic temperatures of Mars the emitted radiation falls mostly in the IR range. Following Wien's law, the maximum of the blackbody spectral radiance for a given temperature is given by $\lambda_{max}$[μm]=2898/$T$[K]. If the maximal and minimal Martian temperatures are $T_{max}$=293K and $T_{min}$=150K then the sensor is designed to work optimally in the range from 9.9μm to 19.3μm.

## 3.1 Atmosphere Transmission Windows

The Martian atmosphere consists mostly of $CO_2$, which has a strongly absorbing band centred at 15μm. The $CO_2$ in the column of air within the cone of view may also act as absorber and emitter (notice that the air is generally at a very different temperature from the ground) around the band. Additionally, water molecules have a very strong absorption at 1.45μm, and a weak absorption at 6.27μm (Martin, 1986).

## 3.2 Reflected Solar Radiance

Since the typical emissivities of Martian soils are different from one, the IR Solar radiation shall be added up to ground emissions (Lienhard and Lienhard, 2006). Assuming that in the IR the Sun radiance on the Martian surface is equal to the one on top (inside atmosphere transmission windows), and that the Martian ground IR reflectivity $r=1-\varepsilon$, with $\varepsilon$ the emissivity, is bounded to 0.1, one can obtain the reflected flux as $E_{reflected}=r \cdot E_{sun}$. The solar flux on Mars surface is, $F_{sun}=E \cdot (R_s/D)^2$ with $R_s$=6.96x$10^8$m the Sun radius, $D$=1.52·1AU= 1.52·1.5x$10^{12}$m the Sun to Mars distance, and where $E$ is the radiance of a blackbody emitting at a temperature T=6000 K (Vázquez *et al.*, 2005).

The measurements must be performed in a range where the ratio of IR radiance emitted by the Martian surface to the solar IR radiance reflected by the Martian surface is significantly greater than one. For instance, figure 3 shows that above 8μm the solar reflected radiance is smaller than 0.5% for the lower ground temperature, $T_g$=150K.

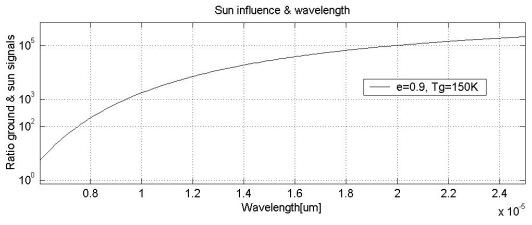



Figure 3: Ratio ground signal/sun radiance vs. wavelength.

Therefore, the GTS channels 8-14μm and 16-20μm are selected taking into account the optimum wavelength range from Wien's law, but trying to avoid the atmospheric absorption bands and the wavelengths in which solar radiation cannot be neglected. Each channel is specialised in the measure of a temperature range where the higher S/N ratio, based on the Planck's law, is achieved (Vázquez *et al.*, 2005). The other GTS band 14.5-15.5μm shall measure the temperature of Martian atmosphere using for that the $CO_2$ emission band.

## 3.3 Reflected Rover Radiance

The MSL rover is a source of error for the GTS, since some parts of it are subjected to temperatures over the ground. These rover elements are mainly the RTG and the rover body, which can reach temperatures 200K and 50K over the atmosphere, respectively. These elements are painted using high emissivity paint, and their temperature shall be recorded on line during Martian operation.

In order to evaluate rover influence as a source of error in the determination of ground temperature it is necessary a thermal conduction and radiation model of Martian environment. In (Lee, 2006) the heating process of ground surface by thermal conduction due to the RTG is studied. The results show a neglected influence in the area focused by the GTS. On the other hand, figure 2 shows a simplified geometrical representation of the environment (GTS thermopiles, rover, atmosphere and Martian surface), from which rover radiance reflected on the ground can be estimated based on a radiation diagram. The radiation model considers that the IR energy reflected by the ground is completely diffuse.

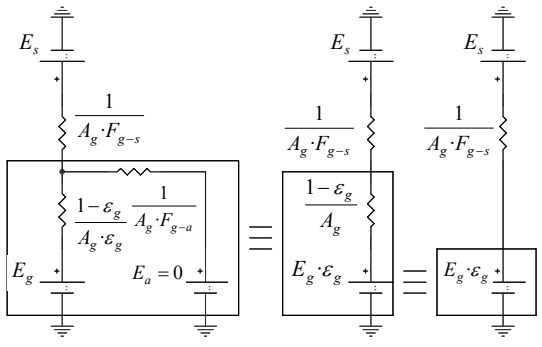



Figure 4: GTS and ground equivalent thermal circuit.

The first step in rover radiance influence analysis considers an ideal situation in which the rover does not exist. The circuit shown in figure 4 represents an electrical analogy of the thermal model, in which voltage generators and currents are equivalent to the energy flux radiated by each body and the heat exchange respectively and resistors represent surface and geometry radiation resistances. The value of the resistors depends on parameters such us the areas and emissivities of the bodies, and the view factor between them (Lienhard and Lienhard, 2006). The atmosphere, as it was said before, is modelled as a transparent body that does not emit energy inside the measurement wavelengths, $E_a=0$. In this way, $A_g$ represents the area of the ground seen by the GTS, $\varepsilon_g$ the emissivity of the ground, $F_{x\text{-}y}$ the view factor between the bodies determined by the subscripts, and finally $E_x$ represents the energy radiated by a blackbody at the temperature of the body determined by the subscript and inside the measurement band of the thermopile. The subscript *g* is for ground, *a* is for atmosphere and *s* is for the thermopiles. The expression of $E_x$ follows Planck's law and takes the form,

$$E_x = \int T(\lambda) \cdot 2hc^2 \Big/ \lambda^5 \cdot \left( e^{hc/\lambda K T_X} - 1 \right) d\lambda \left[ W/m^2 \right] \quad (1)$$

where $T(\lambda)$ is the thermopiles filter transmittance.

Figure 4 shows two successive simplifications of the electrical circuit. The first one obtains an equivalent circuit, assuming that the view factor $F_{g\text{-}a}$ is very close to the unit. This is reasonable because of the small size of the thermopile and environment geometry. The second simplification assumes two things: first the view factor $F_{g\text{-}s}$ is very small and close to zero, since the area of the thermopile compared with the distance between the thermopile and the ground is very small. Second, $\varepsilon_g$ takes real values that go from 0.9 to 1. Thus, the equivalent resistor is dominated by the value of the geometry resistance.

This circuit gives us means for calculating the heat exchange between the environment and the sensor, whose temperature ($T_s$) is known. From it, and based on GTS thermopiles sensibility $G_s$[V/W], the output voltage of the thermopiles (2) can be obtained. This voltage shall be considered as the GTS ideal output, and shall be compared with the real one, once the rover is included in the thermal model. The result of this comparison shall be the error introduced by rover heated bodies. In addition to that, equation (2) depends on the value of ground emissivity, $\varepsilon_g$, which is an *a priori* unknown parameter that introduces also uncertainty in ground temperature determination.

$$V_{out} = G_S \cdot \Delta E \qquad \Delta E = A_g \cdot F_{g-S} \cdot \left( \in_g \cdot E_g - E_S \right) \quad (2)$$

The next step in rover influence analysis includes the rover geometrical and thermal model. Figure 5 stars from a previous simplification of rover body and RTG equivalent circuits. This simplification, whose objective is to obtain the equivalent circuit for these bodies and the atmosphere, is similar to the carried out in figure 4 for Martian ground.

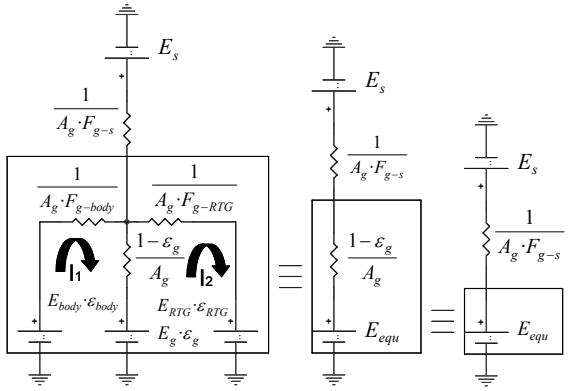



Figure 5: GTS, rover body and ground equivalent thermal circuit.

The equivalent resistance of the squared area of figure 5 can be easily calculated, assuming that view factors between ground and rover body ($F_{g\text{-}body}$) and ground and RTG ($F_{g\text{-}RTG}$) take values close to zero. Thus, the resistor inside the ground branch of the circuit is much smaller than the others, and its value dominates. Equally, the calculation of the equivalent generator of the squared area $E_{equ}$ needs to solve the equations system (3) for $I_1$ and substitute its value in (4). And finally, the second simplification follows the same reasoning of figure 4.

$$\begin{bmatrix} E_g \cdot \in_g - E_{RTG} \cdot \in_{RTG} \\ E_{RTG} \cdot \in_{RTG} - E_{body} \cdot \in_{body} \end{bmatrix} = \begin{bmatrix} \frac{(1-\in_g) F_{g-RTG} + 1}{A_g \cdot F_{g-RTG}} & \frac{-1}{A_g \cdot F_{g-RTG}} \\ \frac{-1}{A_g \cdot F_{g-RTG}} & \frac{F_{g-RTG} + F_{g-body}}{A_g \cdot F_{g-RTG} \cdot F_{g-body}} \end{bmatrix} \begin{bmatrix} I_1 \\ I_2 \end{bmatrix} \quad (3)$$

$$E_{equ} = -I_1 \cdot \frac{(1-\in_g)}{A_g} + E_g \cdot \in_g \quad (4)$$

Newly, equation (4) shows a dependency of ground emissivity, $\varepsilon_g$. In this case temperatures, emissivities and view factors of rover body and RTG appear additionally in the equation as a new source of uncertainty.

# 4 SIMULATIONS ON ROVER AND EMISSIVITY INFLUENCE

The development of practical test with a real or scaled model of the MSL rover is extremely costly, since Martian temperature ranges requires the usage of complicate climatic chambers. From this point of view, this chapter is dedicated to develop preliminary simulations to evaluate or obtain a upper bound of rover and ground emissivity influence in the determination of Martian surface temperature.

The simulations are based on the GTS thermal radiation model described in the previous section. Then, to apply the model, the values of the view factors and the ground area covered by the sensor are required. In order to obtain practical data, simulations using the software package *Thermal Desktop* and a simplified geometrical model of the problem similar to the shown in figure 2, have been carried out. The model assumes that the rover and the ground are in a horizontal plane. The results are shown in table 1. Additionally, it must be pointed out that the value of the energy terms, $E_x$, are obtained based on thermopiles practical data (Sebastián and Gomez-Elvira, 2007).

Table 1: REMS GTS simulation data.

| Body | $\varepsilon$ | $T$ | $A$ | $F_{g\text{-}x}$ |
|---|---|---|---|---|
| GTS | 1 | $T_g$+20K | 1mm$^2$ | 1.99x10$^{-9}$ |
| Ground | 0.91-0.99 | $T_g$ | 99.8m$^2$ | |
| Rover | 0.8 | $T_g$+70K | | 0.0021 |
| RTG | 0.8 | $T_g$+220K | | 0.00101 |

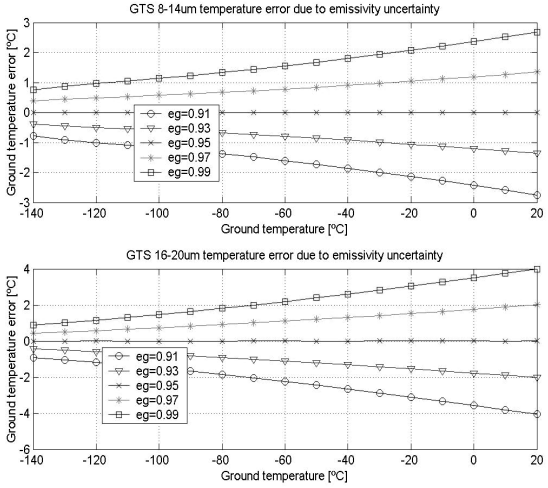



Figure 6: Ground temperature determination error due to ground emissivity uncertainty, supposing an emissivity value of 0.95.

The first simulation, figure 6, tries to analyse the error associated to the determination of ground temperature, generated by the uncertainty in ground emissivity without taking into account rover effects. Temperature error dada are provided for the thermopiles 8-14μm and 16-20μm, considering an *a priory* value for ground emissivity of 0.95. The simulation is run for a possible range of Martian ground temperatures, from 133K to 293K.

The second simulation, figure 7, analyses the error introduced by rover body and the RTG. In this case, ground emissivity is assumed to be known.

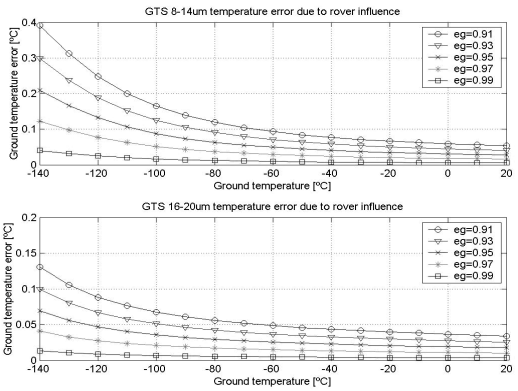



Figure 7: Ground temperature determination error due to rover influence, supposing known values for ground emissivities.

Finally, the last simulation, figure 8, includes the errors associated to ground emissivity uncertainty and rover body and RTG reflections, supposing a value for the ground emissivity of 0.95.

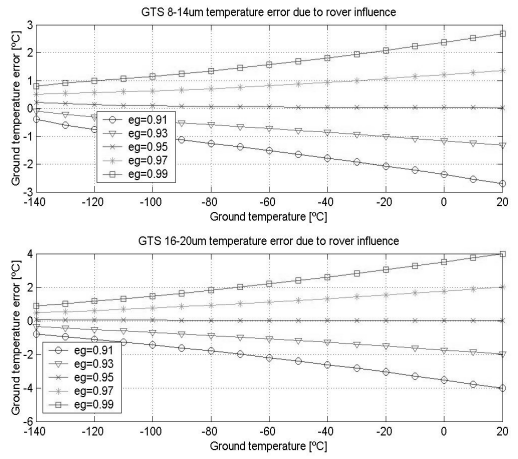



Figure 8: Ground temperature determination error due to rover influence and ground emissivity uncertainty, supposing an emissivity value of 0.95.

# 5 CONCLUSIONS AND FUTURE WORK

The selection of GTS measurement bands is conditioned by Martian atmosphere optical properties, ground temperatures as well as taking into account solar reflected radiance, in order to minimize or neglect the associated errors.

The REMS GTS location and orientation has been selected in order to minimize rover influence, due to the heating process of ground surface by thermal conduction and rover indirect view throughout ground reflected radiance. Additionally, GTS field of view has been maximized in order to increase signal to noise ratio, avoiding rover direct vision.

Simulations on ground emissivity uncertainty have shown an important error contribution in ground temperature determination, reaching error values of ±4K. This error is initially compliant with GTS instrument required accuracy of ±5K, nevertheless is so big that its contribution to the total error budget must be reduced.

A possible solution to deal with this error resorts to study the emissivity of similar soils to those found on Mars. MSL mission includes a set of payload instruments capable of providing detail information about Martian soils composition. Thus, after knowing the king of soil in which the rover is operating, the studied emissivity value can be applied.

Colour pyrometry techniques (Joners and Gardner, 1980) are other possible solution, which could be implemented in order to estimate ground temperature and emissivity at the same time. A possible algorithm consists of four equations (5) with four unknown variables: the emissivities $\varepsilon_g^{8\text{-}14}$ and $\varepsilon_g^{16\text{-}20}$, and ground temperatures $T_{g1}$ and $T_{g2}$. The first two equations are obtained from the equation (2), particularized for the measurement bands (8-14μm, 16-20μm). To complete the four equations a new measurement for a different ground temperature is required, while rover remains still in order to assume constant the value of ground emissivity.

$$\begin{aligned} E_{equ}^{8-14,Tg1} &= f\left(T_{g1},\in_g^{8-14}\right) \quad E_{equ}^{16-20,Tg1} = f\left(T_{g1},\in_g^{16-20}\right) \\ E_{equ}^{8-14,Tg2} &= f\left(T_{g2},\in_g^{8-14}\right) \quad E_{equ}^{16-20,Tg2} = f\left(T_{g2},\in_g^{16-20}\right) \end{aligned} \quad (5)$$

The other source of error studied in this article, rover over temperature effect, generates temperature errors below ±0.4K, for the worst ground temperature conditions. Initially, this error could be neglected in comparison with others, and the

required GTS accuracy. However, this study assumes some simplifications on environment geometry and ground emissivity. For instance, frost formation over the ground or different ground tilts could modify rover reflectance, increasing error contribution.

Therefore, an algorithm to compensate rover influence could be required. The algorithm shall necessarily be based on the model described in section 3, subtracting in (4) the effect of the undesired energy rover terms and solving for $E_g$. In order to do it, an estimation of $\varepsilon_g$, $A_g$, $F_{g\text{-}body}$ and $F_{g\text{-}RTG}$, and the real temperatures of rover bodies are required. The topography of the surface seen by the sensor modifies the view factors between the different environment objects. Thus, in order to carry out this geometrical analysis it is necessary to have a three-dimensional image of rover environment, as well as rover position and orientation. These data shall be provided by NASA. Afterwards, they shall be used to obtain more accurate view factors and areas, or just as a quality control system to confirm the validity of the results. For instance, $CO_2$ frost, shadows with different ground temperature, or extreme ground tilts are different circumstances to be detected.

A more rigorous analysis of rover influence and a possible improvement in this algorithm must include the sensibility of GTS versus the radiation incident angle. Initially, the whole surface area seen by the sensor is weighted equally, this is reasonable for ground emitted radiation since the whole ground is supposed to be at the same temperature. Nevertheless, several small differentials of area, in which the GTS sensibility is different, could be considered instead of a unique ground area. So, IR energy coming from the rover and reflected in these differentials of area must be weighted considering GTS sensibility. Equation 6 shows how the geometrical resistor between ground and RTG would be calculated,

$$\frac{1}{\sum g_i \cdot A_{gi} \cdot F_{g-RTGi}} \tag{6}$$

where $A_{gi}$ is the differential of area $i$, $F_{g\text{-}RTGi}$ is the view factor between the RTG and the differential of area $i$, and $g_i$ is the weighting factor that considers GTS sensibility. Sensibility depends on the incident angle of the radiation and fulfils $A_g = \sum g_i \cdot A_{gi}$.

Finally, the simplify model described in this article must be confirmed using a specialized software such us *Thermal Desktop*, evaluating the global behaviour of the model and not only for obtaining the value of the view factors.

## ACKNOWLEDGEMENTS

The authors would like to express special thanks to all members of the REMS project who are collaborating in the development of the GTS.

## REFERENCES

Joners T.P. Gardner J.L. 1980. Multi-wavelength radiation pyrometry where reflectance is measured to estimate emissivity. *Phys. E: Sci. intrum.*, 1, 3006-319.

Lienhard IV J.H. and Lienhard V. J.H. 2006. *A Heat Transfer Textbook*. web.mit.edu,3rd edition.

Lee C. J. 2006. *MSL study on RTG-to-Ground Interaction Zone*. Applied Sciences Laboratory, Inc. internal report.

Martin T.Z. 1986. Thermal infrared Opacity Of The Mars Atmosphere. *Icarus, 66*, 2-21.

Sebastián E. and Gomez-Elvira J. 2007. Preliminary tests of the REMS GT-sensor, In *ICINCO'07 International Conference on Informatics in Control, Automation and Robotics*, Angers (France).

Vázquez L., Zorzano M.P., Fernández D., McEwan I. 2005. Considerations about the IR Ground Temperature Sensor. *CAB, REMS Technical Note.* Madrid.

# POSTERS

# QOS MULTICAST ROUTING DESIGN USING NEURAL NETWORK

Ming Huang and Shang Ming Zhu
*Department of Computer Science, East China University of Science and Technology, 200237 ShangHai, China*
*huangming@mail.ecust.edu.cn, zhusm@mail.ecust.edu.cn*

Keywords: Quality of service, Hopfield neural network, Multicast routing.

Abstract: In this paper, an algorithm based on Hopfield neural network for solving the problem of determining minimum cost paths to multiple destination nodes satisfying QoS requirements is proposed. The schema of constructing the multicast tree with HNN is emphasized after the analysis of the attributes of the multicast tree. At last, the emulation explores the feasibility of this algorithm.

## 1 INTRODUCTION

With the rapid emergency of the multimedia applications on networks, especially for video conferences, the need for multicast QoS routing mechanism to satisfy the multimedia transmission requirements among conference participants is urgently rising.

Wang and Crowcroft have proved QoS routing with multi-constraint to be a NP-complete problem (Wang et al., 1996). Neural networks have often been formulated to solve NP-complete optimization problems. Tank and Hopfield (Tank et al., 1986) proposed the first neural approach applied in the TSP problem. The advantage of the Hopfield NN is the rapid computational capablility of solving the combinatorial optimization problem. Ahn and Ramakrishna (Ahn et al., 2001) proposed a near-optimal routing algorithm employing a modified Hopfield NN. In this paper, a new Hopfield NN model is proposed to speed up the convergence whilst improving the optimality of the multicast tree constructed under multi-constraint

## 2 PROBLEM FORMULATION

As far as multicast routing is concerned, a network is usually represented as a weighted directed graph $G=(V, E)$, where $V$, the set of vertexes, denotes the set of nodes, and E, the set of directed edges, corresponds to the set of links connecting the nodes.

Suppose the number of nodes is $n$, refer to the $i$th node as $i$ for short, then the adjacency matrix $\mathbf{O}$ describing the initial state of the focused network can be defined as: $\mathbf{O}_{ij}=1$, if the link from $i$ to $j$ exists; elsewise, $\mathbf{O}_{ij}=0$.

We also appoint a node $s \in V$ to denote the source node of a multicast request. The source node acts as the root of the multicast tree to be build which has only out-degree. Correspondingly, let $\boldsymbol{d}=\{d_1,\ldots,d_m\} \subset (V-\{s\})$ denotes the set of the destination nodes with only in-degree. Then the constructed multicast tree, denoted by $T$ and shown in Fig.1, should satisfy:

1) $T$ is a subgraph of $G$;
2) The equal undirected graph $T'$ is a tree;
3) s is the root of $T$';
4) $\boldsymbol{d} \in T$ and every leaf of $T'$ is in $\boldsymbol{d}$;
5) Every vertex in $T$ except s can be accessed from $s$.

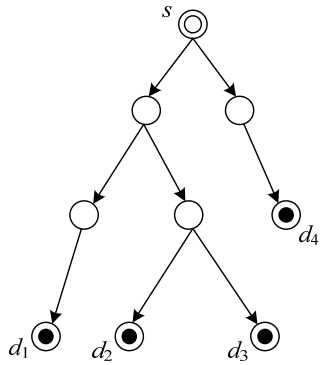



Figure 1: Indicating diagram of multicast tree.

Let the adjacency matrix **A** describing the constructed subgraph *G'* of *G* after multicast routing is defined as: $\mathbf{A}_{ij}$=1, if the link from *i* to *j* is in *G'*; elsewise, $\mathbf{A}_{ij}$=0. Then let's prove the theorems below:

**Theorem 1.** In a multicast tree *T*, there will be one and only one path from *s* to *v* ($v \in T$ and $v \neq s$).

**Proof:** *Existence*. According to attribute 5) of T, *v* can be accessed from *s*, so there is at least one path from from *s* to $d_i$ ($i=1,\ldots,m$);

*Uniqueness*. If there are two different paths from *s* to *v* in *T*, then there will be at least two discoincident sectors composed of directed edges in the two paths. Combined with the joint vertexes, the two sectors form a circle in the equal undirected graph of T. Refer to attribute 2) of multicast tree, the existence of the circle contradict the acyclic attribute of a tree. So there is only one path from *s* to *v* in *T*;

**Theorem 2.** Supposing $s$, $\boldsymbol{d} \in G'$ and the in-degree of *s* is 0, *G'* is a multicast tree T if and only if:

$$\left(\sum_{k=1}^{n}\left(\mathbf{A}^k\right)\right)_{sj} = 1, \text{ where } j \in G' \text{ and } j \neq s \tag{1}$$

$$\prod_{d\in\mathbf{d}}\left(\left(1-\left(\sum_{k=1}^{n}\left(\mathbf{A}^k\right)\right)_{id}\right)^2+\left(\sum_{j\in G'}\mathbf{A}_{dj}\right)^2\right)=0, \tag{2}$$

where $i \notin d$ or $\sum_{j\in G'}\mathbf{A}_{ij} > 0$

**Proof:** *Necessity*. Based on the accessible theory in graph theory, Eq.1 denotes the number of paths from *s* to *j* within a length of *k*. Refer to Theorem 1, Eq.1 is obviously tenable. If $I \in T$, then *i* must locate at a branch of *T* except *i* is a leaf destination node itself. Suppose the branch ends at a leaf $\theta$. According to Theorem 1, there is one and only one path from *s* to $\theta$ and *i* is in the path. Due to attribute 4 of multicast tree, $\theta \in \boldsymbol{d}$, therefore Eq.2 is got.

*Sufficiency*. Obviously, attribute 1 of multicast tree is qualified. Because of Eq.1, attributes 5 is qualified. Considering the in-degree of *s* is 0, *G'* is a tree with *s* as the root, i.e. attributes 2 and 3 is qualified. When Eq.2 is satisfied, for $I \in G'$, except *i* is a leaf destination node itself , there will be at least one leaf $\theta \in \boldsymbol{d}$ which can be accessed from *i* by one and only one path. So attribute 4 is qualified. Therefore, *G'* is a multicast tree.

# 3 NEURAL NETWORK MODEL

The Hopfield NN model for multicast QoS routing, which consists of $n \times n$ neurons connected with each other, is mapped from the corresponding directed graph *G* of the aimed network system with *n* nodes.

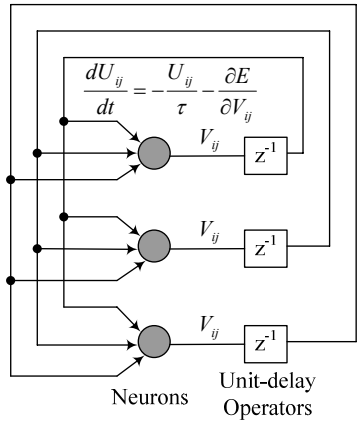



Figure 2: Model of Hopfield neural network.

The output of the neuron at the position (*i*,*j*) is denoted by $V_{ij}$, where $V_{ij}$=1, if the link from *i* to *j* exists; otherwise, $V_{ij}$=0. Obviously, the output matrix $\mathbf{V}=[V_{ij}]_{n\times n}$ is equal to the adjacency matrix **A** of G. Let $U_{ij}$ denotes the input of neuron (*i*,*j*), and define the gain function *g* of the neuron as:

$$\left(\sum_{k=1}^{n}\left(\mathbf{A}^k\right)\right)_{sj} = 1, \text{ where } j \in G' \text{ and } j \neq s \tag{3}$$

$$\frac{dU_{ij}}{dt} = -\frac{U_{ij}}{\tau} - \frac{\partial E}{\partial V_{ij}} \tag{4}$$

Define several link state matrix as: $\mathbf{W}=[W_{ij}]_{n\times n}$, $\mathbf{B}=[B_{ij}]_{n\times n}$, $\mathbf{D}=[D_{ij}]_{n\times n}$ and $\mathbf{L}=[L_{ij}]_{n\times n}$, where $W_{ij}$ is the cost of the link from *i* to *j*, $B_{ij}$ is the bandwidth of the link from *i* to *j*, $D_{ij}$ is the delay of the link from *i* to *j* and $L_{ij}$ is the parcket loss rate of the link from *i* to *j*. The QoS constraints is denoted with $B_w$, $D_w$ and $L_w$ where $B_w$ is the minimal available bandwidth of each selected link, $D_w$ is the maximal available delay of each selected path and $L_w$ is the maximal available packet loss rate of each selected path.

As shown in Fig.2, the dynamic Eq.4 governs the dynamics of the network. The design of the energy function should reflect the attributes of the selected multi-path below:

1) There is no non-existing link in the selected multi-path;
2) There is no input to the source node in the selected multi-path;

3) All destination nodes is in the selected multi-path;
4) According to Theorem 2, the two equations should be satisfied;
5) Satisfy the QoS constraints.
6) The total cost of the selected multi-path must be as low as possible;

To fit the constraints above, we design the suitable energy function as:

$$E = \mu_1 E_1 + \mu_2 E_2 + \mu_3 E_3 + \mu_4 E_4 + \mu_5 E_5 + \mu_6 E_6 \tag{5}$$

$$E_1 = \sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n}\left(\left(V_{ij}O_{ij}-1\right)V_{ij}\right)^2 \tag{6}$$

$$E_2 = \sum_{i=1}^{n}\left(V_{is}\right)^2 \tag{7}$$

$$E_3 = \sum_{d\in\mathbf{d}}\left(\left(\sum_{k=1}^{n}\left(\mathbf{V}^k\right)\right)_{sd}-1\right)^2 \tag{8}$$

$$E_4 = \sum_{\substack{j\in G' \\ j\neq s}}\left(\left(\sum_{k=1}^{n}\left(\mathbf{V}^k\right)\right)_{sj}-1\right)^2 + \sum_{\substack{i\in G' \\ i\neq s \\ i\notin\mathbf{d} \text{ or } \sum_{j\in V}\mathbf{A}_{ij}>0}}\left(\prod_{d\in\mathbf{d}}\left(\left(1-\left(\sum_{k=1}^{n}\left(\mathbf{V}^k\right)\right)_{id}\right)^2 + \left(\sum_{j\in V}\mathbf{V}_{dj}\right)^2\right)\right) \tag{9}$$

$$E_5 = z_1 + H\left(z_2\right) + H\left(z_3\right) \tag{10}$$

$$H\left(z\right) = \int_0^z h\left(z\right)dz,\ h\left(z\right) = \begin{cases} z, \text{ if } z>0 \\ 0, \text{ otherwise} \end{cases} \tag{11}$$

$$z_1 = \sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} H\left(B_{ij}V_{ij}-B_w\right) \tag{12}$$

$$z_2 = \sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} H\left(mD_w-\left(V_{ij}D_{ij}\sum_{k=1}^{n}V_{jk}\right)\right) \tag{13}$$

$$z_3 = \sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} H\left(\left(V_{ij}\left(1-L_{ij}\right)\sum_{k=1}^{n}V_{jk}\right) - m\left(1-L_w\right)\right) \tag{14}$$

$$E_6 = \sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n}\left(W_{ij}V_{ij}\right)^2 \tag{15}$$

Obviously, $E_1$, $E_2$ and $E_3$ refer to constraints 1, 2 and 3 respectively. According to Theorem 2, when $E_4$ get minimal (i.e. zero), the constraint 4 is attained. $E_5$ represents the integration of QoS constraints on bandwidth, delay and loss rate. As $z_1$, $z_2$ and $z_3$ represents the deviation of the QoS constraints and the minus value is meaningless, so we filter the minus value with an integral computation in $E_5$. With $z_1$, $z_2$ and $z_3$, conditions of links near root will influence the deviation more severely.

# 4 EMULATION

Fig.3 shows the topology of the delegating network system. The source node which promotes the routing request and four destination nodes have already been signed out in the graph. Each link in the network is labeled with a link status vector consists of cost $W$, bandwidth $B$(MB), delay $D$(ms) and loss rate $L$. The QoS constraints are: $B_w$=2MB, $D_w$=8ms, $L_w$=$10^{-3}$.

In this case, we set coefficient $\mu_1$ =80000, $\mu_2$ =40000, $\mu_3$=160000, $\mu_4$ =500, $\mu_5$ =400, $\mu_6$ =400, $\lambda$=1, and $\Delta t = 2\times10^{-8}$. By emulation, we could obtain the global optimal solution shown in Fig.3.

# 5 CONCLUSIONS

In this paper, we propose a new multicast tree selection algorithm to simultaneously optimize multiple QoS parameters which is based on Hopfield neural network. The result of emulation shows that the utilization of Hopfield neural network is an available method to solve QoS routing problems. Furthermore, by the massive parallel computation of neural network, it can find near optimal route quickly, so the real-time requirement of routing in high-speed network system could be satisfied.

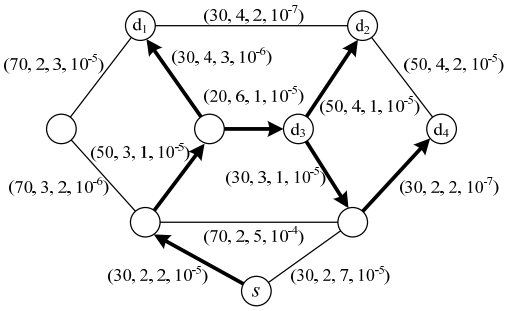



Figure 3: Topology of network system with parameters and selected multi-path.

## REFERENCES

Wang Z, Crowcroft J, 1996. *Quality of service for supporting multimedia applications. 14th. IEEE JSAC.*

Tank, D.W., Hopfield, J.J, 1986. *Simple neural optimization networks: an A/D converter, signal decision network, and a linear programming circuit. 33th. IEEE Transactions on Circuits and Systems.*

Ahn, C.W., Ramakrishna, R.S., Kang, C.G., Choi, I.C., 2001. *Shortest Path Routing Algorithm Using Hopfield Neural Network. 37th. Electronics Letters.*

B.M. Waxman, 1988. *Routing of multipoint connections. Vol.9. IEEE TSAC.*

S.Chen, K.Nahrstedt, Y.Shavitt, 2000. *A QoS-aware multicast routing protocol. Vol.18. IEEE Journal on Selected Areas in Communications*.

Q.Sun, 1999. *A genetic algorithm for delay-constrained minimumcost multicasting. Technical Report. IBR. TU Braunschweig.* Braunschweig. Germany.

# CONTROL THEORETIC APPROACH TO ANALYSIS OF RANDOM BRANCHING WALK MODELS ARISING IN MOLECULAR BIOLOGY

Andrzej Swierniak
*Silesian University of Technology, Department of Automatic Control, Akademicka 16, Gliwice, Poland*
*andrzej.swierniak@polsl.pl*

Keywords: Systems modelling, random branching walks, control applications in biology.

Abstract: We present two models of molecular processes described by infinite systems of first order differential equations. These models result from branching random walk processes used to represent the evolution of particles in these problems. Using asymptotic techniques based on Laplace transforms it is possible to characterize the asymptotic behavior of telomeres shortening which is supposed to be the mechanism of aging and evolution of cancer cells with increasing number of copies genes responsible for coding causing drug removal or metabolisation. The analysis in both cases is possible because they could be represented by systems with positive feedbacks.

## 1 PROBLEM STATEMENT

Shortening of telomeres is one of the supposed mechanisms of cellular aging and death. The hypothesis is that each time a cell divides it loses pieces of its chromosome ends. These ends are called telomeres and consist of repeated sequences of nucleotides, telomere units. When a critical number of telomere units is lost, the cell stops dividing. Telomeres are assumed to consist of telomere units repeated sequences of nucleotides. When a chromosome replicates each newly synthesized strand loses one telomere unit at one of its ends. This means that the pair of daughter chromosomes each has one old unchanged strand and one new, one unit shorter. Once a critical number of telomere units is lost a so called Hayflick checkpoint is reached and the cell stops dividing. Under this assumption, only the length of the shortest telomere will matter and thus a chromosome is said to be of type j if its shortest telomere has j remaining units (Arino, Kimmel, Webb, 1995).
The amount of DNA per cell remains constant from one generation to another because during each cell cycle the entire content of DNA is duplicated and then at each mitotic cell division the DNA is evenly apportioned to two daughter cells. However, recent experimental evidence shows that for a fraction of DNA, its amount per cell and its structure undergo continuous change. Gene amplification can be enhanced by conditions that interfere with DNA synthesis and is increased in some mutant and tumor cells. Increased number of gene copies may produce an increased amount of gene products and, in tumor cells, confer resistance to chemotherapeutic drugs. Amplification of oncogenes has been observed in many human tumor cells and also may confer a growth advantage on cells which overproduce the oncogene products (for an overview see e.g. survey in (Stark, 1993)).
We present models of this two phenomena using branching random walk machinery. The asymptotic properties of them could be found using methods of Laplace transforms and spectral analysis. Conclusions resulting from this analysis are general because we demonstrate that the models could be represented by the linear systems with positive feedbacks and therefore we are able to use some well known results from standard control theory of infinite dimensional control systems.

## 2 MODEL OF TELOMERE SHORTENING

The simplest model of telomere shortening is due to Levy et al.(1992). It is based on the following assumptions:

1. Each chromosome consists of 2 strands: upper and lower, each of them having 2 endings right and left.
2. Number of telomere units on both endings may be written as quadruple ($a$, $b$; $c$, $d$), where $a$ and $c$ correspond to left and right ending of the upper strand, while $b$ and $d$ correspond to left and right ending of the lower one. The only possible combinations are of the form ($n$–1, $n$; $m$, $m$) or ($n$, $n$; $m$,$m$–1).
3. Cells having chromosomes described by a quadruple ($n$–1, $n$; $m$, $m$) while dividing result in progenies of types ($n$–1, $n$–1; $m$, $m$–1) and ($n$–1, $n$; $m$, $m$). The similar rule takes place for the second type leading to the situation in which one of the progenies is always of the same type as the parent cell while the other is missing two sequences each on a different ending of a different strand.
4. The process ends when telomere endings are short enough; without loss of generality it may be viewed as the case ($n$–1, $n$; 0, 0) or (0, 0; $m$, $m$–1). In this case the cell does not divide and the single progeny is identical with the parent.

The transformation takes the form:

$$(n-1,n;m,m) \quad \begin{cases} \rightarrow (n-1,n;m,m) \\ \rightarrow (n-1,n-1;m,m-1) \end{cases} \tag{1}$$

$$(n,n;m,m-1) \quad \begin{cases} \rightarrow (n,n;m,m-1) \\ \rightarrow (n-1,n;m-1,m-1) \end{cases} \tag{2}$$

$$(n-1,n;0,0) \rightarrow (n-1,n;0,0) \tag{3}$$

$$(0,0;m,m-1) \rightarrow (0,0;m,m-1) \tag{4}$$

We can observe that such "two-dimensional" process may be simplified by introducing indices $k$ and $l$ denoting total number of units on both upper and lower strand for left and right endings respectively.
Denoting:

$$k = \begin{cases} 2n & \text{if} \quad (n,n;m,m-1) \quad \text{appears} \\ 2n-1 & \text{if} \quad (n-1,n;m,m) \quad \text{appears} \end{cases} \tag{5}$$

$$l = \begin{cases} 2m & \text{if} \quad (n-1,n;m,m) \quad \text{appears} \\ 2m-1 & \text{if} \quad (n,n;m,m-1) \quad \text{appears} \end{cases} \tag{6}$$

the feasible transformations are as follows:

$$(k,l) \quad \begin{cases} \rightarrow (k,l) \\ \rightarrow (k-1,l-1) \end{cases} \tag{7}$$

$$(k,0) \rightarrow (k,0) \tag{8}$$

$$(0,l) \rightarrow (0,l) \tag{9}$$

Defining $i = \min(k, l)$ leads to the simplest form of the admissible transitions:

$$i \quad \begin{cases} \rightarrow i \\ \rightarrow i-1 \end{cases} \tag{10}$$

and

$$0 \rightarrow 0 \tag{11}$$

Index $i$ describing the state of the cell in the sense of the telomere's length may be called the type of the cell. Dynamics of this model could be represented by a system of state different equations the asymptotic behavior of which has a polynomial form as a function of the number of generation.
Deterministic model treats all cells as homogeneous, not taking into account its variability dealing mainly with different life time. The simplest approaching to real world is to treat cell doubling times as random variables with exponential distribution characterized by the same parameter $\alpha$. The evolution process becomes a branching random walk with an expected number of cells of type $j$ originated of the ancestor of type $i$ denoted by $N_{ij}(t)$ given by the following state equation:

$$\dot{N}_{ij}(t) = \alpha N_{ij+1}(t), i \geq j \geq 0 \tag{12}$$

For finite number of nonzero initial conditions:

$$N_i(0) > 0, \ \text{i} \leq M \tag{13}$$

we have:

$$N_j(t) = \sum_{i=j}^{M} \frac{\alpha t^{i-j}}{(i-j)!} N_i(0) \tag{14}$$

where $N_j(t)$ is an average number of cells in the state $j$.
Once more the solution (exact solution and not only asymptotic expansion as it has been the case in the previously discussed discrete model) has a form of polynomial function of time. Moreover if we assume that the random variables representing doubling time has an arbitrary distribution the same in each generation the asymptotic formula for the average number of cells in all states could be also given by (14) with the parameter of exponential distribution substituted by the inverse of the average doubling time resulting from the assumed distribution.

We demonstrate that these rather strange asymptotic characteristics and the generality of their form is related to the positive feedback which could be discovered in all the three models of telomere shortenings.

# 3 MODEL OF GENE AMPLIFICATION

We consider a population of neoplastic cells stratified into subpopulations of cells of different types, labeled by numbers $i = 0, 1, 2, \ldots$. If the biological process considered is gene amplification, then cells of different types are identified with different numbers of copies of the drug resistance gene and differing levels of resistance. Cells of type 0, with no copies of the gene, are sensitive to the cytostatic agent. Due to the mutational event the sensitive cell of type 0 can acquire a copy of gene that makes it resistant to the agent. Likewise, the division of resistant cells can result in the change of the number of gene copies. The resistant subpopulation consists of cells of types $i = 1, 2, \ldots$.
The probability of mutational event in a sensitive cell is of several orders smaller than the probability of the change in number of gene copies in a resistant cell. Since we do not limit the number of gene copies per cell, the number of different cell types is denumerably infinite.
The hypotheses are as follows:

1. The lifespans of all cells are independent exponentially distributed random variables with means $1/\lambda_i$ for cells of type $i$.
2. A cell of type $i \geq 1$ may mutate in a short time interval $(t, t+dt)$ into a type $i+1$ cell with probability $b_i\,dt + o(dt)$ and into type $i-1$ cell with probability $d_i\,dt + o(dt)$.
3. A cell of type $i = 0$ may mutate in a short time interval $(t, t+dt)$ into a type 1 cell with probability $\alpha\,dt + o(dt)$, where $\alpha$ is several orders of magnitude smaller than any of $b_i$s or $d_i$s, i.e.
$$\alpha << \min(d_i, b_i), \quad i \geq 1. \qquad (15)$$
4. The chemotherapeutic agent affects cells of different types differently. It is assumed that its action results in fraction $u_i$ of ineffective divisions in cells of type $i$.
5. The process is initiated at time $t = 0$ by a population of cells of different types.

The mathematical model has the following form:

$$\begin{cases} \dot{N}_0(t) = [1-2u(t)]\,\lambda N_0(t) - \alpha N_0(t) + dN_1(t) \\ \dot{N}_1(t) = \lambda N_1(t) - (b+d)N_1(t) + dN_2(t) + \alpha N_0(t) \\ \ldots \\ \dot{N}_i(t) = \lambda N_i(t) - (b+d)N_i(t) + dN_{i+1}(t) + bN_{i-1}(t), \\ \qquad\qquad\qquad\qquad\qquad\qquad i \geq 2 \\ \ldots \end{cases} \qquad (16)$$

where $N_i(t)$ denotes the expected number of cells of type $i$ at time $t$, and we assume the simplest case, in which the resistant cells are insensitive to drug's action, and there are no differences between parameters of cells of different type ($b_i = b > 0$, $d_i = d > 0$, $\lambda_i = \lambda > 0$, $u_i = 0$, $i \geq 1$, $\lambda_0 = \lambda$, $u_0 = u$).
The first step in the analysis is to evaluate the fate of the drug resistant subpopulation without a constant inflow from the drug sensitive subpopulation. In other words we assume that it is possible to destroy completely the sensitive subpopulation and we are interested only how the drug resistant cancer cells will develop. The analysis can be limited in this case to equations with $i \geq 1$. The asymptotic behavior of the DNA repeats may be analyzed using inverse Laplace transforms and asymptotic formulae for integration of special functions for the case where the initial condition contained only one nonzero element $N_1(0) = 1$, while $N_i(0) = 0$, $i > 1$. It is possible to extend that approach to the case of two or more non-zero elements. The solution decays exponentially to zero in this case, as $t \to \infty$ for:

$$d > 0, b > 0, \lambda > 0, d > b, \qquad (17)$$

$$\sqrt{d} - \sqrt{b} > \sqrt{\lambda} \qquad (18)$$

To analyze the conditions under which it is possible to eradicate the tumor or in other words to ensure that the entire tumor population converges to zero we may represent the model (16) in the form of the closed-loop system with two components. One part of this system is infinite dimensional and linear and represents the drug resistant subpopulation. The second part of the system is given by the first bilinear equation of the model and describes behavior of the drug sensitive subpopulation. The model may be viewed as a system with positive feedback stability of which may be analyzed using generalized Nyquist type criterion (Swierniak, *et al.* , 1999) in the case when we assume a constant therapy protocol. The analysis for other protocols could be also performed using more sophisticated tools of stability analysis.
In the similar way we may consider more general models of anticancer therapy under evolving drug

resistance such as a multi-drug chemotherapy, models including phase specificity in the sensitive compartment or models which take into account partial sensitivity of some neoplastic subpopulations (Swierniak, Smieja, 2005).

# 4 CONCLUSION REMARKS

In this paper we have studied asymptotic properties of two models of molecular processes each of them modeled by the random branching walk models. The properties of these models are strictly related with their structure which when considered from system theoretic point of view includes always the positive feedback. Moreover although the models have the form of infinite dimensional state equations linear or bilinear the asymptotic analysis may be performed rigorously using control theoretic tools resulting from the closed loop structure of these models. Yet another molecular process which could be analyzed using similar techniques is the evolution of tandem repeats in microsatellite DNA Once more random branching walk could be used as a basis for the model construction. Nevertheless in this case there is no positive feedback which has been used by us to simplify the asymptotic analysis of the two processes considered in this paper.

# REFERENCES

Arino O., Kimmel M., Webb G.F. 1995. Mathematical modeling of telemore sequences, *J. Theoretical Biology,* v.177, 45-57.

Kimmel M., Swierniak A., Polanski A., 1998. Infinite-dimensional model of evolution of drug resistance of cancer cells, *J. Mathematical Systems, Estimation, and Control*, v.8, 1-16.

Levy M.Z., Allstrop R.C., Futchert A.B., Grieder C.W., Harley C.B. , 1992. Telomere end-replication problem and cell aging, *J. Molec. Biol*., v.225, 951-960.

Stark G.R. , 1993. Regulation and mechanisms of mammalian gene amplification, *Adv. Cancer Res*., v. 61, 87-113.

Swierniak A., Polanski A., Kimmel M., Bobrowski A., Smieja J. , 1999. Qualitative analysis of controlled drug resistance model - inverse Laplace and semigroup approach, *Control and Cybernetics*, v.28, 61-74.

Swierniak A., Smieja J. , 2005. Analysis and optimization of drug resistant and phase-specific cancer chemotherapy models, *Math.Biosciences and Engineering,* v. 2, 657-670

# ON THE SAMPLING PERIOD IN FUZZY CONTROL ALGORITHMS FOR SERVODRIVES

## *A Strategy for Variable Sampling*

Dan Mihai
*University of Craiova, Decebal Blvd, 107, Craiova, Romania*
*dmihai@em.ucv.ro*

Keywords: Fuzzy control, Adaptive sampling period, On-line timing, Microcontroller, Servodrives.

Abstract: The paper deals with a variable control sampling period for the fuzzy control algorithms implemented on low inertia servodrives. The robustness of the fuzzy control strategy is extended on the sampling period values and hence an adaptive sampling algorithm is proposed. The author analyzes the possibility to vary continuously the sampling frequency upon a basic process variable. Principles, models and simulation results inserted here give reliance in this technique and an enhancement of the fuzzy control implementation. The distribution of the sampling moments in different adaptive conditions and the behaviour of the servodrive are obtained by means of some models and simulations in accordance with the real-time target hardware system.

# 1 INTRODUCTION

The author found (Mihai, 2001) that the fuzzy logic has the ability to drive the system in good conditions for very different control sampling period - T values over more than a magnitude order. In such a context, the idea to vary continuously the T value, in accordance with a dynamic parameter of the system, finds a suitable application area. The standard digital models become non-linear because the variable coefficients and the classical algorithms are very sensitive to T. A variable T means, in almost all the approaches, acquisitions with a variable frequency. Less studies and experience concern the real-time control with a variable cycle. Most of the involved authors and equipment use several pre-computed constant values T. Computer graphics applications refer to an adaptive sampling in term of an adjustment of the sampling resolution in exploiting the image (Adamson, 2005). The adapting sampling in the fuzzy control could also provide means to reduce noise in computer graphics, like for global illumination algorithms (Xu, 2006). Also some other special or non-conventional application fields implement an adaptive sampling (radio telemetric system for missiles, drying processes in food industry). Although some papers still prove the natural idea that a sampled-data fuzzy controller recovers the performance of the continuous-time fuzzy controller as the sampling period approaches zero (Do Wan, 2007), several authors have noticed that the fuzzy control is flexible and reliable for a low rate control sampling (Popescu, 1997; Mihai, 2001). Using an adaptive sampling frequency for the control of a servodrive is a complex task because of the fast reaction speed of such a system and its high associated performance.

# 2 THE FUZZY CONTROLLER AND T VALUES

Although T seems, apparently, not being an essential variable for the main characteristics of a FLC (fuzzy sets and the rule base), this parameter is involved in a fuzzy loop in two ways:
- as a real - time “integration step” of the system, (acquisition–processing–control cycle;
- as an input FLC variables generator by:

$$Va_k = V(kT);\ \Delta Vb_k = V_k - V_{k-1};\ Vck = (V_k - V_{k-1}) / T \quad (1)$$

The author considered a low inertia servodrive with DC motors. The figure 1 gives the essential structure for the drive with disk rotor motor and an encoder. The FLC entries are the normalized position error and the normalized variation of the position error:

$$\varepsilon_{\alpha n\,k} = \varepsilon_{\alpha n\,k} \cdot (\alpha^* - \alpha_k) / \alpha^* \quad (2)$$

$$\Delta\varepsilon_{\alpha nk} = \varepsilon_{\alpha nk} - \varepsilon_{\alpha nk-1} = [T \cdot (-\omega_k)]_n = \omega_k \cdot \Delta\varepsilon_{\alpha nk\,max} / \Omega_{max} \quad (3)$$

$\alpha^*/\alpha_k$ - the position set point/ the actual position; $N_\alpha^*/N_{\alpha k}$ – same, in encoder pulses; $\Delta N_{\alpha k}$ - pulses encoder during T; $c_k$, $c_{kout}$ - the computed control and its outputted value; $Norm_i$: normalization blocks; CPB: Control Processing Block; PS - Power supply; $T_{gen}$ - torque generator; M - motor; En - encoder. The encoder has $N_{p/r}$ pulses per revolution and the speed is computed with:

$$\omega_{ks} \approx \frac{\alpha_k - \alpha_{k-1}}{T} = \frac{2\pi \cdot k_{div} \cdot \Delta N_k}{N_{p/r} \cdot T} = c_{sp} \cdot \Delta N_k \quad (4)$$

The simulation results from figure 2 are obtained using a FLC from Fuzzy Toolbox (Guley, 1995), with fuzzy sets and rules presented in (Mihai, 2008). The quality of the results is proved by the final position error (null) and the fuzzy state-space trajectory, between the initial point (10, 0) and the final point (0,0)–last window. When T increases, the FLC task becomes more difficult. Although for the whole range the controller succeeds in bringing the system to the final point, some internal ringing or steps appear. It is obviously also that for low sampling frequencies, the speed (quite well filtered by the mechanical system) is far from the position error variation. Another model is designed as a fuzzy position / speed loop for the same system but with a Look-Up-Table (LUT) FLC-figure 3. An additional argument for the adaptation of T is given by the real-time recordings presented in figure 4, for an on-line inference fuzzy control (Mihai, 2006). During every T (SPER), each falling edge of the encoder pulses (PULSE) leads to a fast hardware interrupt routine (INTO) that up-dates the FLC entries. FUZZ is the fuzzyfier task, INFER-the on-line inference task, DEFUZ-defuzzification task and AUX concerns other processing tasks, like savings. The 2 diagrams were recorded for different conditions, revealing the ability of the FLC to manage the microcontroller resources even at maximum speed, when the processing algorithm is interrupted at maximum rate. However, the available time is very depending on the motor speed. A higher speed could lead to the situation when the control processor is no more able to fulfil the real-time task inside T. Its adaptation to the speed would be the solution. For adding robustness related with the load variation, a special strategy was proposed by the author keeping the same reference LUT. Additional procedures were implemented for on-line adaptation of the control to the load value, both by an estimated current and some external computations and decisions blocs.

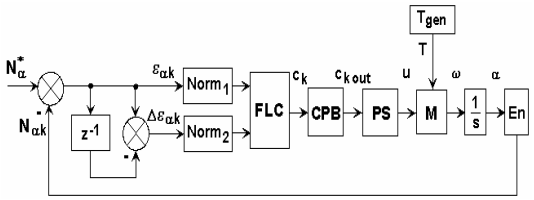



Figure 1: The servodrive with FLC.

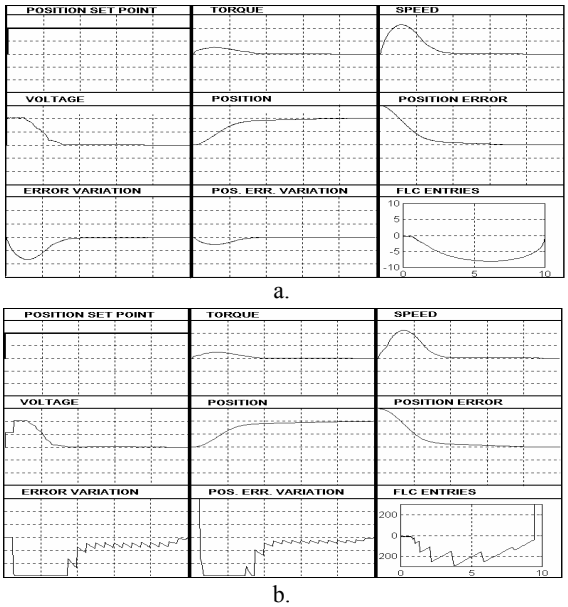



Figure 2: Results: T=2.456 ms (a) and T=50 ms (b).

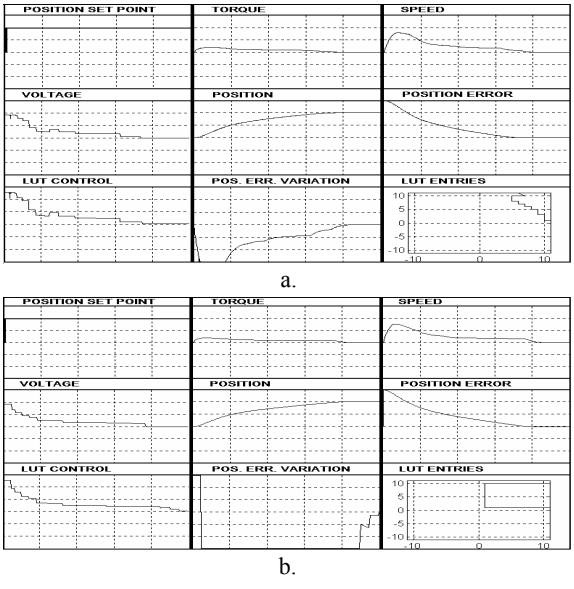
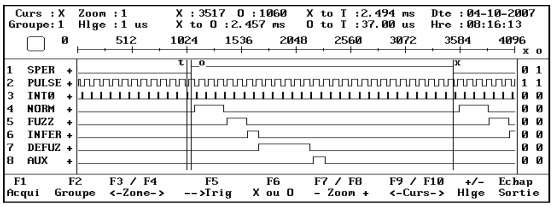
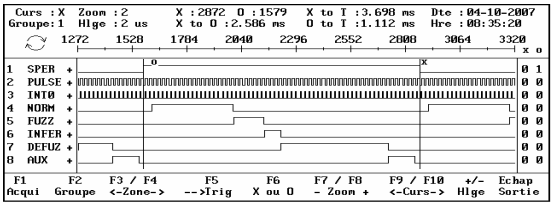



Figure 3: Results for T = 2.456 ms (a) and T= 50 ms (b) with a LUT based FLC.




a.




b.

Figure 4: An on-line T in fuzzy control.

# 3 A FUZZY CONTROLLER WITH ADAPTIVE T VALUES

The idea is to relate T with the variation rate of the main variable of the system. During the intervals with small variations (or in steady state regimes), T is greater and during the high rate dynamic regimes, T decreases. The variation for the (generic) variable v from the step t induces an adequate adaptation of T at t+δt. The figure 5 gives an image of the principle and helps for obtaining some relations. A first possibility is to evaluate the amplitude variation for the main variable during a constant time interval (easily in real-time). Another idea is to use an amplitude quantization of the v variable using a constant step Δ and to evaluate then the time intervals associated with this variation. They can be directly assimilated with the adapted T. Next relations make connects the derivative value and Δ.

$$v = f(t);\ \Delta = f(t_{i+1}) - f(t_i)\ ;\ \mathrm{tg}\,\alpha_i = \Delta / T_i \qquad (5)$$

$$T_i = \Delta / \mathrm{tg}\,\alpha_i \cong \Delta / ( \left| |df / dt|_{ti} \right| + \varepsilon) \qquad (6)$$

Δ could be chosen by practical considerations. If ***f*** is known, (5) is useful for evaluate T. If not, $t_i$ result by detecting the amplitude thresholds and by that the next T value is available. ε is for a limitation of max.T. A limitation is also necessary for minT, (systemic, on-line processing constraints):

$$T_{min} \le T \le \Delta / \varepsilon \qquad (7)$$

For a servodrive where the main variable is the position, let be the speed ω the variable v (the variation of the position). It is more suitable to use a T adaptation in accordance with the variation rate not after the amplitude of the v variable. Indeed, for that last case, even in a steady-state regime, the sampling rate is high and for a low speed during a strong dynamic regime the sampling has a slow rate. If the main characteristic variable of the system is the speed, v could be the acceleration. The next idea is to adapt also the step value Δ upon another characteristic variable of the controlled system. The results for two Δ (constant) values are depicted by figure 6, with the distribution of the sampling moment. Position 1 is the sampled position with an adaptive period. "State space FLC path" is the trajectory of the system. The global behaviour is good for both variants, the final position error being null and the system response in speed and position being a smooth one. Figures 7a and 7b give the variations of T for Δ=2.456 and Δ=20, during the whole regime. The max/min rate values are almost 100/Δ=10 and 35/Δ=50. Figures 7c, 7d make visible the large variation range of the max/min T along 3 magnitude ranges (logarithmic scale).

The next idea is to use a variable quantization step for adapting T as a double adaptive sampling strategy. Another adaptation parameter is involved. The fig. 8 gives the elements for that, considering the speed as an additional modulator (by its change rate) for the adaptive sampler of the position. In this way, it is no more necessary to make different experiments in order to adopt the best step value Δ (Q). The distribution of the sampling moments is different (fig. 8a) and the step value is variable (fig. 8b). The image of the new sampled position is given by fig. 8c. The results from 8d concern another values range for the modulator of the adaptive sampler (a larger one – see Q_ad). It is depicted also the quantified speed – Speed 1, as the source of the modulator for the quantization step necessary for the sampler with double adaptive T. The overshoot for the position is related with its quantized final values.

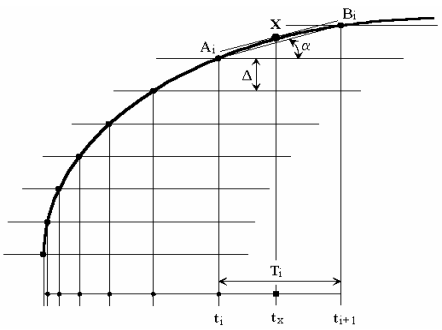



Figure 5: For adaptation of T values.

# 4 CONCLUSIONS

A variable sampling frequency could give a better control. This approach leads to some serious robustness problems for the classical algorithms but not for the fuzzy control. A good robustness regarding the sampling period for the fuzzy control induced the idea to try a control with adaptive sampling period. This idea is applied for a servodrive - a fast and precise system. Several variants were considered: adaptive sampler with a constant quantized step, with a multi-step modulation and with a continuous variation.

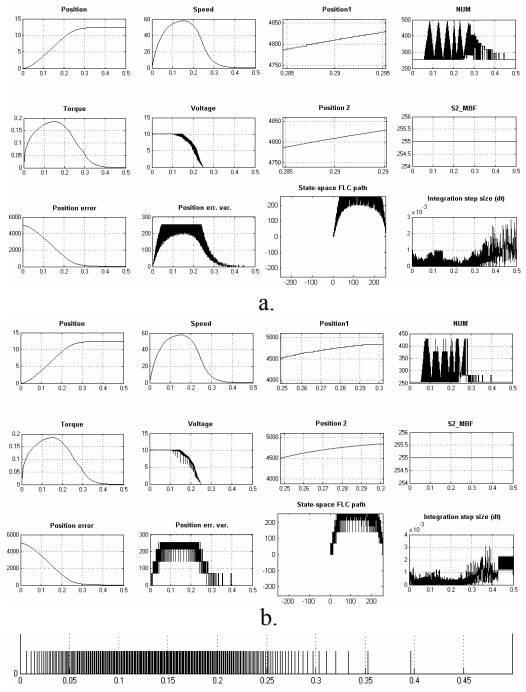



Figure 6: Main variables of the system for a step $\Delta = 2.5$ (a), $\Delta = 20$ (b) and T evolution.

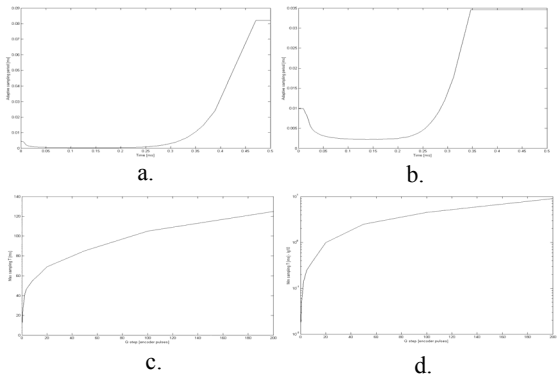



Figure 7: T evolution in time and upon $\Delta$.

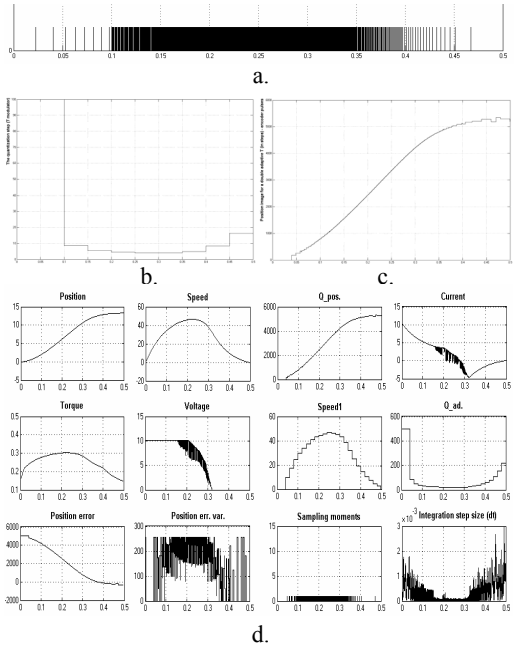



Figure 8: Adaptive FLC /variable step.

# REFERENCES

Adamson, A., Alexa, M., A. Nealen, 2005. Adaptive sampling of intersectable models exploiting image and object-space coherence, *Symp. on Interactive 3D graphics and games I3D '05*, ACM Press.

Do Wan, K., Jin Bae, P., Young Hoon, J., 2007. Effective digital implementation of fuzzy control systems based on approximate discrete-time models, *Automatica (IFAC Jour.)*, Volume 43 , Issue 10, pp. 1671-1683.

Gulley N., Jang,J.S. R., 1995 *Fuzzy Logic Toolbox for use with Matlab*, The Mathworks, Inc.

Mihai, D., 2001. Robustness of the Fuzzy Digital Control Regarding the Sampling Frequency for a Servodrive System, *ELECTROMOTION '01*, Bologna, pp. 431-435.

Mihai, D., 2006. An Optimized Fuzzy Control Algorithm for Servodrives. Some Real-Time Experiments, *IS '06*, London Proc., pp. 192-197.

Mihai, D., 2008. On the Sampling Period in Standard and Fuzzy Control Algorithms for Servodrives. A multicriterial design and a timing strategy for constant sampling, *INICO 08*, Funchal - Madeira, Portugal.

Popescu, G.S., Pastravanu, A., Bogdan, I., 1997. PWM AC/DC converter with reduced sampling frequency fuzzy control, *Proc. of the IEEE International Sym. on Industrial Electronics*, Vol. 3, pp. 1228 – 1231.

Xu, Q., Xing, L., Wang, W., Sbert M., 2006. Adaptive sampling based on fuzzy inference, *Proc. of the 4th int. conference on Computer graphics and interactive techniques in Australasia and Southeast Asia* GRAPHITE, ACM Press.

# SYNTHESIS OF THE LOW-PASS AND HIGH-PASS WAVE DIGITAL FILTERS

B. Psenicka, F. Garcia-Ugalde and A. Romero Mier y Teran
*Universidad Nacional Autonoma de México, Mexico*
*pseboh@servidor.unam.mx, rommiermn@hotmail.com*

Keywords: Wave Digital Filter, Algorithm, Implementation in DSP.

Abstract: In this paper we propose a very simple procedure for the design and analysis of low-pass and high-pass wave digital filters derived from reference filter given in the lattice configuration. Wave Digital Filters derived from reference filter in lattice configuration can be designed with excellent pass band properties. They can be proposed and implemented without the knowledge of classical filter theory. In this paper we present tables for Butterworth, Chebychev and Elliptic low-pass filter design. In the examples we demonstrate programs in MATLAB that permits analyze the attenuation properties of the designed filters. In the end of our article we realize wave digital filter using Embedded Target for Texas instruments TMS320C6000 DSP Platform. The model of the WDF was created by means of serial and parallel blocks that were added to the window Simulink Library Browser between common Used Blocks.

## 1 INTRODUCTION

Wave digital lattice filters are derived from LC-filters. The reference filter consists of parallel and serial connections of several elements. Since the load resistance $R_L$ is not arbitrary but dependent on the element or source to which the port belongs, we cannot simply interconnect the elements to a network. The elements of the filters are connected with the assistance of serial and parallel adapters. These adapters in the discrete form are connected in one port by delay elements. The possibility of changing the port resistance can be achieved using parallel and serial adapters. These adapters contain the necessary adders, multipliers and inverters. In this paper we use adapters with three ports. The block of the serial and parallel adapter and theirs signal-flow diagram are shown in figures 1 and 2. (Fettweis, 1973)

The coefficient of the 3-port reflection-free serial adapter in figure 1A) is calculated from the port resistances $R_i$ i=1,2 by equation (1). (Fettweis and Meerkoetter, 1975)

$$B = \frac{R_1}{R_1 + R_2} \tag{1}$$

The coefficient of the reflection-free parallel adapter in figure 1B) can be calculated from the port conductances $G_i$ i=1,2 by eq. (2).

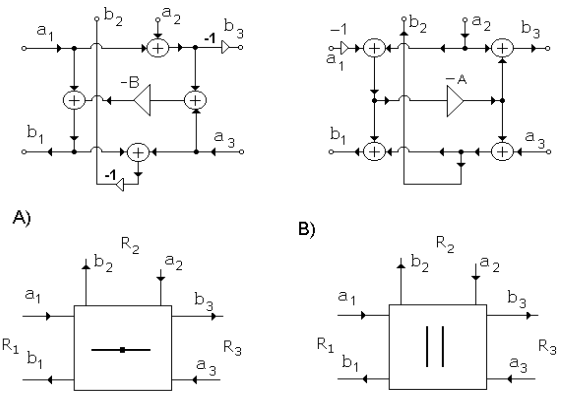



Figure 1: A) Three port serial adapter whose port 3 is reflection-free and its signal flow-graph, B) Three port parallel adapter whose port 3 is reflection-free and its signal flow-graph.

$$A = \frac{G_1}{G_1 + G_2} \tag{2}$$

The coefficients of the dependent parallel adapter in the figure 2B) can be get from the port conductances $G_i$ i=1,2,3 by eq. (3)

$$A_1 = \frac{2G_1}{G_1 + G_2 + G_3} \qquad A_2 = \frac{2G_2}{G_1 + G_2 + G_3} \tag{3}$$

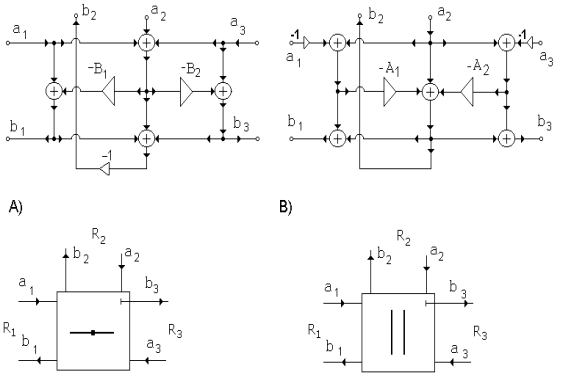



Figure 2: A) Three port serial dependent adapter and its signal flow-graph, B) Three port parallel dependent adapter and its signal flow-graph.

The coefficient of the dependent serial adapter in the figure 2A) can be obtained from the port resistances $R_i$ i=1,2,3 by (4)

$$B_1 = \frac{2R_1}{R_1+R_2+R_3} \qquad B_2 = \frac{2R_2}{R_1+R_2+R_3} \quad (4)$$

When connecting adapters, the network must not contain any feedback loops without a delay element in order to guarantee that the structure is realizable. This means that we cannot connect the dependent adapters from figure 2A) and 2B). The three-port dependent parallel adapter is reflection free at port 3 if $G_3 = G_1 + G_2$ and three port dependent serial adapter is reflection free if $R_3 = R_1 + R_2$.

# 2 EXAMPLES

In this part we shall demonstrate in the examples calculation of the low-pass and high-pass wave digital filter. The most important components for the realization of wave digital filters according to the Fettweis procedure are the ladder LC filters. The tables for wave digital structures was designed for the corner frequency $f_1 = 1/(2\cdot\pi) = 0.159155$ and sampling frequency $f_s = 0.5$

## 2.1 Realization of the Low-pass Filter

In the first example we shall realize Butterworth low-pass of the $5^{th}$ order and $A_{max} = 3dB$. In the figure 3 we show the structure of a $5^{th}$ order ladder LC reference Butterworth filter and its corresponding block connection in the digital form.

First we must calculate from FIG. 3 wave port resistances $R_1, R_2, R_3, R_4$, and finally the coefficients of the

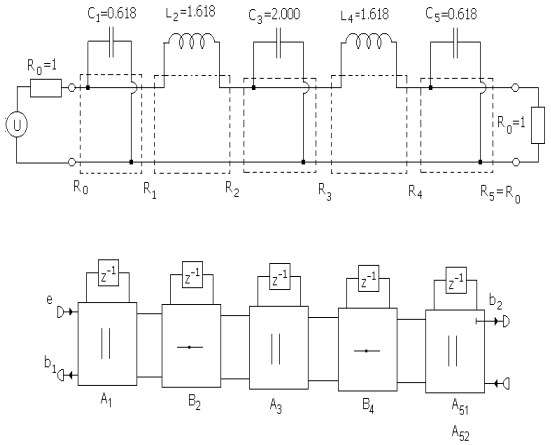



Figure 3: LC reference Butterworth low-pass filter and its corresponding block connection.

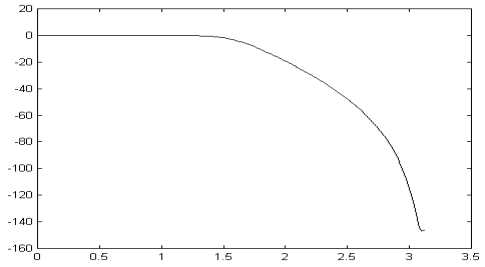


Figure 4: Frequency response of the Butterworth low-pass filter.

parallel and serial adapters $A_1, B_2, A_3, B_4$ and the coefficients of the dependent parallel adapter $A_{51}, A_{52}$ according to equations (1)-(3).

$$G_1 = G_0 + C_1 = 1.618 \qquad G_3 = G_2 + C_3 = 2.447$$

$$R_1 = 1/G_1 = 0.618 \qquad R_3 = 1/G_3 = 0.408$$

$$R_2 = R_1 + L_2 = 2.236 \qquad R_4 = R_3 + L_4 = 2.026$$

$$G_2 = 1/R_2 = 0.447 \qquad G_4 = 1/R_4 = 0.493$$

$$A_1 = \frac{G_0}{G_1+C_1} = 0.618 \qquad B_4 = \frac{R_3}{R_3+L_4} = 0.201$$

$$B_2 = \frac{R_1}{R_1+L_2} = 0.276 \qquad A_{51} = \frac{2G_4}{G_4+C_5+G_5} = 0.443$$

$$A_3 = \frac{G_2}{G_2+C_3} = 0.182 \qquad A_{52} = \frac{2C_5}{G_4+C_5+G_5} = 0.556$$

The program for the analysis of the wave digital filter of the $5^{th.}$ order written in MATLAB follows, and the frequency response obtained by MATLAB is presented in the figure 4. The equation in the program for computing XN1-XN4, BN4-BN1, N1-N9 and YN(i) was received from the structure in the Fig. 5

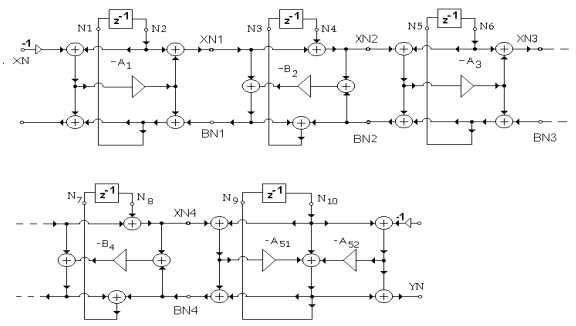



Figure 5: Butterworth wave digital low-pass filter of the $5^{th}$ order.

```
A1=0.618146; B2=0.276524; A3=0.182858;
B4=0.201756; A51=0.467553; A52=0.947276;
N2=0; N4=0; N6=0; N8=0; N10=0; XN=1;
for i=1:1:50
   XN1=A1*XN-A1*N2+N2;   XN2=XN1+N4;
   XN3=-A3*XN2-A3*N6+N6;   XN4=XN3+N8;
   BN4=XN4-A51*XN4+2*N10-A51*N10-A52*N10;
   BN3=XN3-B4*XN4-B4*BN4;
   BN2=XN2-A3*XN2+BN3+N6-A3*N6;
   BN1=XN1-B2*XN2-B2*BN2;
   N1=XN*A1-A1*N2+BN1;  N3=BN1+BN2;
   N5=-A3*XN2-A3*N6+BN3;  N7=BN3+BN4;
   N9=-A51*XN4+N10-A51*N10-A52*N10;
   YN(i)=-A51*XN4+2*N10-A51*N10-A52*N10
   N2=N1; N4=N3; N6=N5; N8=N7; N10=N9; XN=0;
end
[h,w]=freqz(YN,1,50)
   plot(w,20*log10(abs(h)))
```

## 2.2 Design of the Low-pass and High-pass Chebychev Filter

In the second example we shall propose low-pass and high-pass Chebychev WDF for n=5, $A_{max} = 3$ dB. From the table 4 we get the values of the WDF $A_1 = 0.223$, $B_2 = 0.226$, $A_3 = 0.182$, $B_4 = 0.192$, $A_{5,1} = 0.383$ and $A_{5,2} = 0.360$. Using previous MATLAB program we obtain the attenuation of the Chebychev filter presented in the Fig. 6. High-pass we get by changing in the preceding program N2=-N1, N4=-N3 N6=-N5, N8=-N7 and N10=-N9.

## 2.3 Realization of the Low-pass Cauer WDF

In the figure 7 the structure of the $5^{th}$ order ladder LC reference Cauer low-pass filter is shown. The values of the LC filter was obtained from the table C 0550 for $\Theta = 30^o$ for $A_{max} = 1.2494$ dB, $A_{min} = 70.5$ dB and $\Omega_s = 2.0000$ (Saal, 1979). Parallel resonant LC circuits in the low-pass filter Fig. 7 will be realized

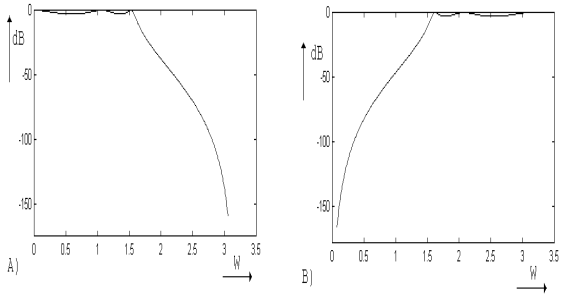



Figure 6: Frequency response of the low-pass and high-pass Chebychev filter.

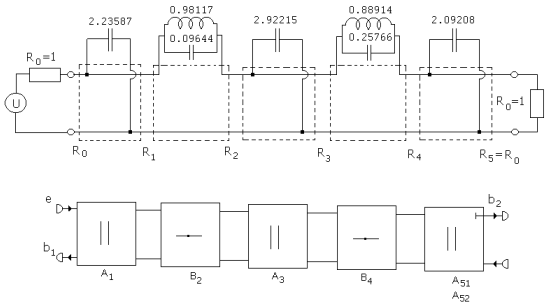



Figure 7: LC reference Cauer low-pass filter.

by digital structure according the Fig. 8.

With the assistance of the following MATLAB program we can calculate coefficients of the WDF $A_1$, $B_2$, $A_3$, $B_4$, $A_{51}$ and $A_{52}$. The attenuation of the low-pass Cauer filter is presented in Fig. 9. The program was obtained from the structure in the Fig. 10. The input data of the LC filter was obtained from the catalog of the Cauer filter (Saal, 1979).

```
C(1)=2.235878; C(3)=2.922148; C(5)=2.092084;
L(2)=0.981174; L(4)=0.889139; C(2)=0.096443;
```

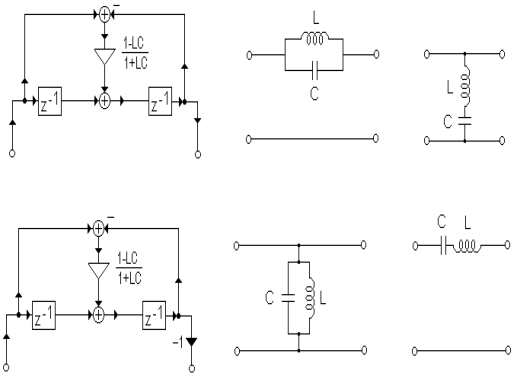



Figure 8: Discrete realization of parallel and serial LC circuits.

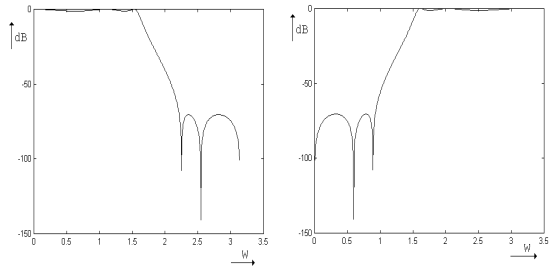


Figure 9: Frequency response of the Cauer low-pass and high-pass filter.

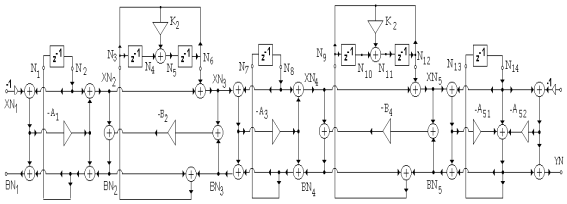


Figure 10: Structure of the Cauer low-pass filter n=5.

```
C(4)=0.257662; XN=1; N2=0; N4=0; N6=0; N6=0;
N8=0; N10=0; N12=0; N14=0; G(1)=1;
G(2)=G(1)+C(1); R(2)=1/G(2);
R(3)=R(2)+1/(C(2)+1/L(2));
G(3)=1/R(3); G(4)=G(3)+C(3); R(4)=1/G(4);
R(5)=R(4)+1/(C(4)+1/L(4)); G(5)=1/R(5);
K(2)=(L(2)*C(2)-1)/(L(2)*C(2)+1);
K(4)=(L(4)*C(4)-1)/(L(4)*C(4)+1);
A(1)=G(1)/G(2); B(2)=R(2)/R(3);
A(3)=G(3)/(G(3)+C(3));
B(4)=R(4)/R(5); A(5,1)=2*G(5)/(G(5)+C(5)+1);
A(5,2)=2/(G(5)+C(5)+1);
for i=1:1:500
    XN1=A(1)*XN+N2-A(1)*N2; XN2=XN1+N6;
    XN3=-A(3)*XN2+N8-A(3)*N8; XN4=XN3+N12;
    BN4=XN4-XN4*A(5,1)+2*N14-N14*A(5,1)-A(5,2)*N14;
    BN3=-BN4*B(4)+XN3-B(4)*XN4;
    BN2=XN2-A(3)*XN2+BN3+N8-A(3)*N8;
    BN1=XN1-XN2*B(2)-BN2*B(2);
    N1=XN*A(1)-N2*A(1)+BN1; N3=BN1+BN2;
    N5=-K(2)*N3+K(2)*N6+N4; N7=-XN2*A(3)+BN3-N8*A(3);
    N9=BN3+BN4; N11=N10-N9*K(4)+N12*K(4);
    N13=-A(5,1)*XN4+N14-A(5,1)*N14-A(5,2)*N14;
    YN(i)=-A(5,1)*XN4+2*N14-A(5,2)*N14-A(5,1)*N14;
    N2=N1; N4=N3; N6=N5; N8=N7; N10=N9; N12=N11;
    N14=N13; XN=0;
end
[h,w]=freqz(YN,1,500); plot(w,20*log10(abs(h)))
```

High-pass can be obtained by changing in program N2=-N1, N4=-N3 N6=-N5, N8=-N7, N10=-N9, N12=-N11, N14=-N13. In the Fig. 9 the attenuation of low-pass and high-pass filter are presented.

## 2.4 Realization of Wave Digital Filters in DSP C6711 by Simulink

The simulink model showed in Figure 11 corresponds to the realization of a wave digital filter application on TMS320C6711 DSK using Embedded Target for Texas Instruments TMS320C6000 DSP Platform. The model of the WDF was created by means of serial and parallel block that were added to the window Simulink Library Browser between Commonly Used Blocks. In the input and output of the WDF were added ADC and DAC convertes of the TMS320C6711 that are in the window Simulink Library Browser, Embedded Target for TI C6000 DSP and C6711DSK Board Support. This model created in Code Composer Studio project can be see in Figure 13 and can run on the DSP C6711.

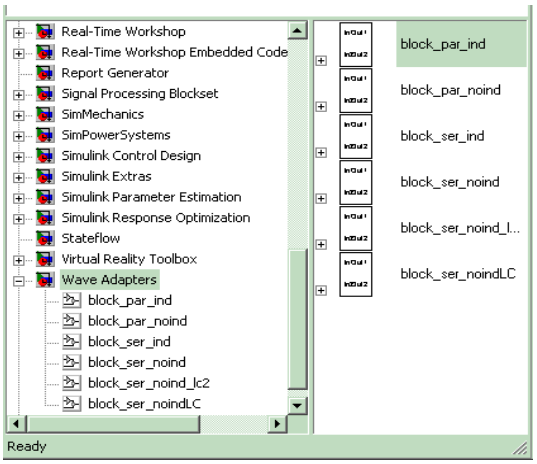



Figure 11: Realization of Wave Digital Filter in TMS320C6711 by Simulink.

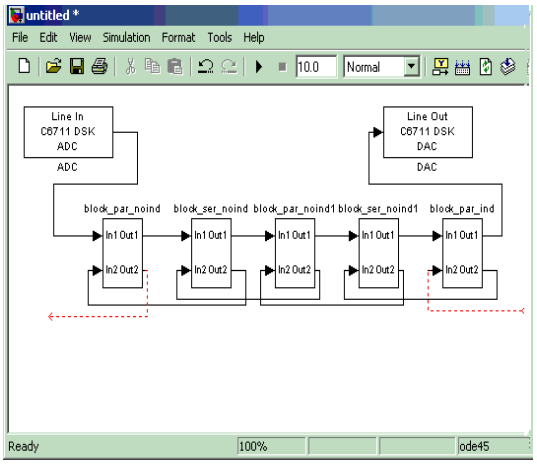



Figure 12: Realization of Wave Digital Filter in TMS320C6711 by Simulink.

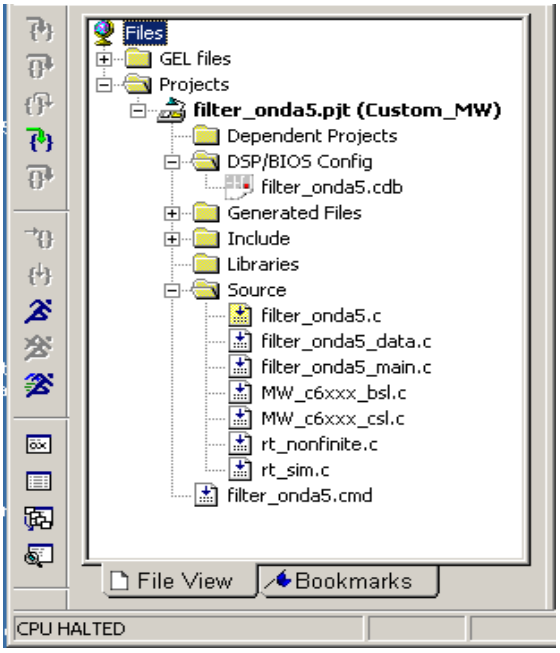



Figure 13: Realization of Wave Digital Filter in TMS320C6711 by Simulink.

## 3 CONCLUSIONS

Though the structure of the wave digital filter is more complicated than other structures, the algorithm for implementation on the DSP is very simple and it is very easy to propose the general algorithm for the arbitrary order of the filter. These structures are not so sensitive to the error of quantization as other types of the filters. With small modification of the presented programs can be created another tables of the WDF. The parts of the programs can be utilized for implementing of the wave digital filters in digital signal processors DSP.

## REFERENCES

Fettweis, A. (1973). Digital filter structures related to classical filter networks. In *Arch. Electron. Uebertragungstechnik*. AEU.

Fettweis, A. and Meerkoetter, K. (1975). On adaptors for wave digital filters. In *IEEE Trans. on Accoustics, Speech and Signal Processing*. IEEE.

Saal, R. (1979). Handbuch zum filterentwurf. In *AEG Telefunken*. AEG Telefunken.

# APPENDIX

## Tables of the Wave Digital Filters

### Tables of the Butterworth Wave Digital Filters

In the Fig. 14 is the tolerance scheme of the Butterworth filter and in the Fig. 15 is the structure of the Butterworth Wave Digital Filter (WDF).

The structure is created by cascade connection of the parallel and serial adapters. If at the begin is parallel reflection free adapter, at the end of the structure in case n odd must be connected parallel dependent adapter in order to realize load resistance $R_L$. In case n even at the end we have to connect serial dependent adapter.

In the table 2 are the elements of the Butterworth WDF for various attenuation $A_{max}$ in the passband. $A_i$ and $B_i$ are the coefficients of the parallel and serial adapters respectively.

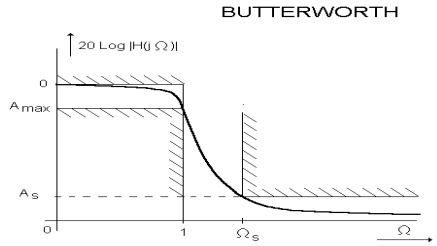



Figure 14: Attenuation of Butterworth WDF.

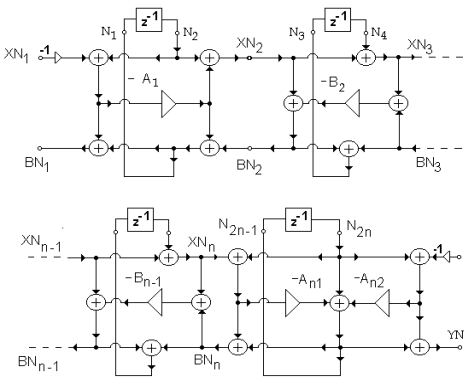



Figure 15: Digital structure of Butterworth WDF.

### Tables of the Chebychev Wave Digital Filter

In Fig. 16 is the tolerance scheme and the attenuation of the the Chebychev filter. The structure of the Chebychev wave digital filters is demonstrated in the Fig. 17. This structure is created by connection of

Table 1: Elements of Butterworth WDF n=3, $A_{max}$ in [dB].

| $A_{max}$ | $A_1$ | $B_2$ | $A_{31}$ | $A_{32}$ |
|---|---|---|---|---|
| 0.1 | 0.6519 | 0.3790 | 0.5497 | 0.9454 |
| 0.2 | 0.6248 | 0.3422 | 0.5099 | 0.9310 |
| 0.3 | 0.6083 | 0.3209 | 0.4858 | 0.9211 |
| 0.4 | 0.5964 | 0.3059 | 0.4685 | 0.9134 |
| 0.5 | 0.5864 | 0.2943 | 0.4548 | 0.9069 |
| 0.6 | 0.5791 | 0.2849 | 0.4434 | 0.9014 |
| 0.7 | 0.5723 | 0.2769 | 0.4337 | 0.8964 |
| 0.8 | 0.5664 | 0.2700 | 0.4252 | 0.8920 |
| 0.9 | 0.5611 | 0.2639 | 0.4176 | 0.8878 |
| 1.0 | 0.5562 | 0.2585 | 0.4208 | 0.8840 |

Table 2: Elements of Butterworth WDF n=5, $A_{max}$ in [dB].

| $A_{max}$ | $A_1$ | $B_2$ | $A_3$ | $B_4$ | $A_{51}$ | $A_{52}$ |
|---|---|---|---|---|---|---|
| 0.1 | 0.702 | 0.387 | 0.286 | 0.318 | 0.602 | 0.981 |
| 0.2 | 0.687 | 0.365 | 0.265 | 0.295 | 0.578 | 0.972 |
| 0.3 | 0.678 | 0.353 | 0.253 | 0.281 | 0.563 | 0.974 |
| 0.4 | 0.671 | 0.344 | 0.244 | 0.271 | 0.553 | 0.971 |
| 0.5 | 0.666 | 0.337 | 0.238 | 0.264 | 0.544 | 0.969 |
| 0.6 | 0.662 | 0.331 | 0.232 | 0.258 | 0.537 | 0.968 |
| 0.7 | 0.658 | 0.326 | 0.228 | 0.252 | 0.531 | 0.966 |
| 0.8 | 0.655 | 0.322 | 0.224 | 0.248 | 0.526 | 0.965 |
| 0.9 | 0.652 | 0.318 | 0.220 | 0.244 | 0.521 | 0.964 |
| 1.0 | 0.649 | 0.314 | 0.217 | 0.240 | 0.517 | 0.962 |

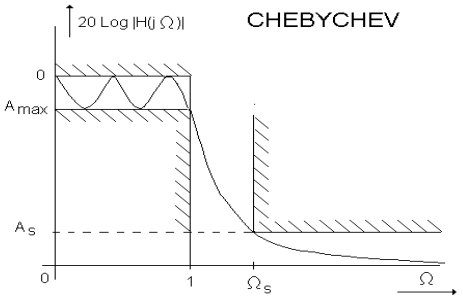



Figure 16: Attenuation of Chebychev WDF.

the parallel and serial adapters terminated at the port 3 with delay element. At the end of the structure must be connected parallel dependent adapter to realize for n odd load resistance $R_L = 1$.
In table 4 are the elements of the Chebychev WDF for various attenuation $A_{max}$ in the passband and order filter n=5. The tables was designed for sampling frequency $f_s = 0.5$.

**Tables of the Cauer Wave Digital Filters**

In the Fig 18 is the attenuation of the Cauer WDF. The structure of the Cauer WDF is presented in the Fig 19. LC parallel resonant circuits in longitudinal branch are realized by serial adapters connected in the $3^{rd}$ port with two delay elements. The filter with n odd

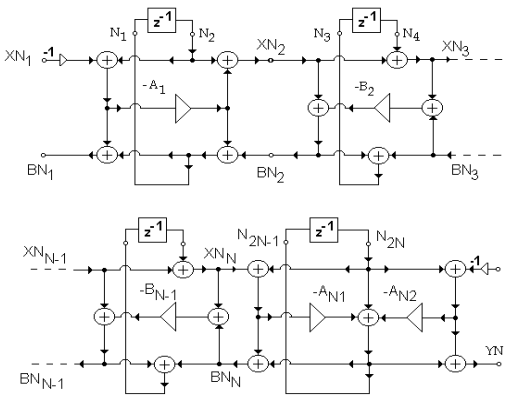



Figure 17: Structure of Chebychev WDF filter.

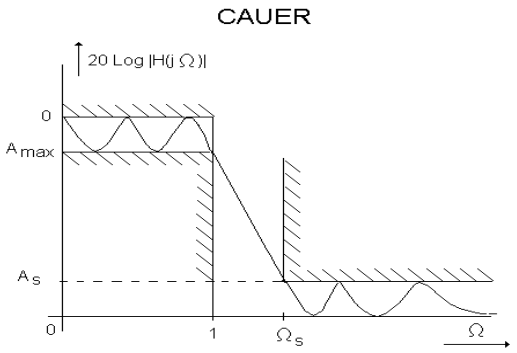



Figure 18: Attenuation of Cauer WDF.

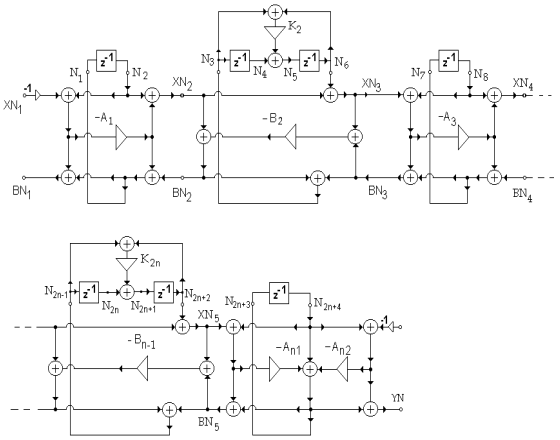



Figure 19: Digital structure of Cauer WDF.

must be terminated with parallel dependent adapter.

In the tables 5 and 6 are the elements of the Cauer WDF for various attenuation $A_{max}$ in passband and sampling frequency $f_s = 0.5$. In the next section we shall demonstrate in the examples calculation of the Wave digital low-pass and high-pas filters.

Table 3: Elements of Chebychev WDF n=3, $A_{max}$in [dB].

| $A_{max}$ | $A_1$ | $B_2$ | $A_{31}$ | $A_{32}$ |
|---|---|---|---|---|
| 0.10 | 0.4922 | 0.3022 | 0.4620 | 0.7574 |
| 0.25 | 0.4341 | 0.2774 | 0.4310 | 0.6812 |
| 0.50 | 0.3852 | 0.2599 | 0.4126 | 0.6114 |
| 1.00 | 0.3307 | 0.2496 | 0.3995 | 0.5293 |
| 2.00 | 0.2695 | 0.2445 | 0.3929 | 0.4331 |
| 3.00 | 0.2300 | 0.2442 | 0.3925 | 0.3696 |

Table 4: Elements of Chebychev WDF n=5, $A_{max}$ in [dB].

| $A_{max}$ | $A_1$ | $B_2$ | $A_3$ | $B_4$ | $A_{51}$ | $A_{52}$ |
|---|---|---|---|---|---|---|
| 0.10 | 0.465 | 0.253 | 0.216 | 0.224 | 0.417 | 0.737 |
| 0.25 | 0.419 | 0.240 | 0.205 | 0.213 | 0.398 | 0.672 |
| 0.50 | 0.369 | 0.231 | 0.197 | 0.204 | 0.386 | 0.596 |
| 1.00 | 0.319 | 0.226 | 0.191 | 0.198 | 0.379 | 0.516 |
| 2.00 | 0.261 | 0.225 | 0.185 | 0.193 | 0.379 | 0.422 |
| 3.00 | 0.223 | 0.226 | 0.182 | 0.192 | 0.383 | 0.360 |

Table 5: Elements of Cauer WDF n=3, $\Omega_s = 4.8097$, $K_2 = -0.93686$.

| $A_m$ | $A_s$ | $A_1$ | $B_2$ | $A_{31}$ | $A_{32}$ |
|---|---|---|---|---|---|
| 0.0004 | 24.7 | 0.7504 | 0.5766 | 0.7314 | 0.9629 |
| 0.0017 | 30.7 | 0.7209 | 0.4907 | 0.6664 | 0.9374 |
| 0.0039 | 34.3 | 0.6673 | 0.4569 | 0.6272 | 0.9160 |
| 0.0109 | 38.7 | 0.6190 | 0.4063 | 0.5778 | 0.8803 |
| 0.0279 | 42.8 | 0.5706 | 0.3612 | 0.5339 | 0.8365 |
| 0.0436 | 44.8 | 0.5461 | 0.3459 | 0.5140 | 0.8114 |
| 0.0988 | 48.3 | 0.4983 | 0.3161 | 0.4804 | 0.7573 |
| 0.1773 | 50.9 | 0.4614 | 0.2979 | 0.4590 | 0.7110 |
| 0.2803 | 53 | 0.4306 | 0.2855 | 0.4442 | 0.6699 |
| 1.2494 | 60 | 0.3147 | 0.2593 | 0.4118 | 0.4999 |

Table 6: Elements of Cauer WDF n=5, $\Omega_s = 2.0000$ $K_2 = -0.827$ $K_4 = -0.627$.

| $A_{max}$ | $A_s$ | $A_1$ | $B_2$ | $A_3$ | $B_4$ | $A_{51}$ | $A_{52}$ |
|---|---|---|---|---|---|---|---|
| 0.0017 | 41.3 | 0.657 | 0.399 | 0.330 | 0.436 | 0.734 | 0.915 |
| 0.0039 | 44.8 | 0.627 | 0.373 | 0.311 | 0.404 | 0.690 | 0.893 |
| 0.0109 | 49.2 | 0.585 | 0.343 | 0.290 | 0.369 | 0.639 | 0.856 |
| 0.0279 | 53.3 | 0.543 | 0.318 | 0.272 | 0.341 | 0.596 | 0.813 |
| 0.0510 | 55.3 | 0.522 | 0.306 | 0.264 | 0.323 | 0.577 | 0.788 |
| 0.0988 | 58.8 | 0.479 | 0.288 | 0.251 | 0.310 | 0.546 | 0.736 |
| 0.1773 | 61.4 | 0.446 | 0.277 | 0.243 | 0.298 | 0.527 | 0.691 |
| 0.2803 | 63.5 | 0.418 | 0.270 | 0.237 | 0.289 | 0.514 | 0.652 |
| 1.2494 | 70.5 | 0.309 | 0.256 | 0.221 | 0.269 | 0.492 | 0.487 |

# ROTATION-INVARIANT IRIS RECOGNITION
## *Boosting 1D Spatial-Domain Signatures to 2D*

Stefan Matschitsch, Herbert Stögner, Martin Tschinder
*School of Telematics & Network Engineering, Carinthia Tech Institute, Austria*

Andreas Uhl
*Department of Computer Sciences, University of Salzburg, Austria*
*andreas.uhl@sbg.ac.at*

Keywords: Biometric authentication, iris recognition, rotation invariance.

Abstract: An iris recognition algorithm based on 1D spatial domain signatures is improved by extending template data from mean vectors to 2D histogram information. EER and shape of the FAR curve is clearly improved as compared to the original algorithm, while rotation invariance and the low computational demand is maintained. The employment of the proposed scheme remains limited to the similarity ranking scenario due to its overall FAR/FRR behaviour.

## 1 INTRODUCTION

With the increasing usage of biometric systems in general the interest in non-mainstream modalities rises naturally. Iris recognition systems are claimed to be among the most secure modalities exhibiting practically 0% FAR and low FRR which makes them interesting candidates for high security application scenarios. An interesting fact is that the iris recognition market is strongly dominated by Iridian Inc. based technology which is based on algorithms by J. Daugman (Daugman, 2004). The corresponding feature extraction algorithm employes 2D Gabor functions. However, apart from this approach, a wide variety of other iris recognition algorithms has been proposed in literature, most of which are based on a feature extraction stage involving some sort of transform (see e.g. (Ma et al., 2004; Zhu et al., 2000) for two examples using a wavelet transform).

Controlling the computational demand in biometric systems is important, especially in distributed scenarios with weak and low-power sensor devices. Integral transforms (like those already mentioned or others like DFT, DCT, etc.) cause substantial complexity in the feature extraction stage, therefore feature extraction techniques operating in the spatial domain have been designed (e.g. (Ko et al., 2007)) thus avoiding the additional transform complexity.

An additional issue causing undesired increase in complexity is the requirement to compensate for the possible effects of eye tilt. For example, the matching stage of the Daugman scheme involves multiple matching stages using several shifted versions of the template data which is a typical approach. As a consequence, rotation invariant iris features are highly desired to avoid these additional computations.

Global iris histograms (Ives et al., 2004) combine both advantages, i.e. rotation invariant features extracted in the spatial domain thus providing low overall computational complexity. However, FAR and FRR are worse compared to state of the art techniques. A recent approach (Du et al., 2006) uses rotation invariant 1D signatures with radial locality extracted from the spatial domain. Still, also the latter technique suffers from unsatisfactory FAR and FRR and thus is only recommended to be used in a similarity ranking scheme (i.e. determining the $n$ closest matches). In this work we aim at improving this algorithm.

In Section 2, we will review the original version of the algorithm and then describe the improvements conducted. Section 3 provides experimental results. We first describe the experimental settings (employed data and software used). Subsequently, we present and discuss our experimental results providing EER improvements over the original version of the algorithm. Section 4 concludes the paper and gives outlook to future work.

# 2 ROTATION INVARIANT IRIS SIGNATURES

Iris texture is first converted into a polar iris image which is a rectangular image containing iris texture represented in a polar coordinate system. Note that the ISO/IEC 19794-6 standard defines two types of iris imagery: rectilinear images (i.e. images of the entire eye like those contained in the CASIA database) and polar images (which are basically the result of iris detection and segmentation). As a further preprocessing stage, we compute local texture patterns (LTP) from the iris texture as described in (Du et al., 2006). We define two windows $T(X,Y)$ and $B(x,y)$ with $X > x$ and $Y > y$ (we use $15 \times 7$ pixels for $T$ and $9 \times 3$ pixels for $B$). Let $mT$ be the average gray value of the pixels in window $T$. The LTP value of pixels in window $B$ at position $(i, j)$ is then defined as

$$LTP_{i,j} = |I_{i,j} - mT|$$

where $I_{i,j}$ is the intensity of the pixel at position $(i, j)$ in $B$. Note that due to the polar nature of the iris texture, there is no need to define a border handling strategy. LTP represents thus the local deviation from the mean in a larger neighbourhood.

In order to cope with non-iris data contained in the iris texture, LTP values are set to non-iris in case 40% of the pixels in $B$ or 60% of the pixels in $T$ are known to be non-iris pixels.

## 2.1 The Original 1D Case

The original algorithm (Du et al., 2006) computes the mean of the LTP values of each row (line) of the polar iris image and concatenates those mean values into a 1D signature which serves as the iris template. Clearly, this vector is rotation invariant since the mean over the rows (lines) is not at all affected by eye tilt. If more then 65% of the LTP values in a row are non-iris, this signature element is ignored in the distance computation. In order to assess the distance between two signatures, the Du measure is suggested (Du et al., 2006) which we apply in all variants.

## 2.2 The 2D Extension

LTP row mean and variance capture first order statistics of the LTP histogram. In order to capture more properties of the iris texture without losing rotation invariance we propose to employ the row-based LTP histograms themselves as features (since histograms are known to be rotation invariant as well and have been used in iris recognition before (Ives et al., 2004)). This adds a second dimension to the signatures of course (where the first dimension is the number of rows in the polar iris image and the second dimension is the number of bins used to represent the LTP histograms).

In fact, we have a sort of multi-biometrics-situation resulting from these 2D signatures, since each histogram could be used as a feature vector on its own. We suggest two fusion strategies for our 2D signatures:

1. Concatenated histograms: the histograms are simply concatenated into a large feature vector. The Du measure is applied as it is in the original version of the algorithm.
2. Accumulated errors: we compute the Du measure for each row (i.e. each single histogram) and accumulate the distances for all rows.

The iris data close to the pupil are often said to be more distinctive as compared to "outer" data. Therefore we propose to apply a weighting factor $> 1$ to the most "inner" row, a factor $= 1$ to the "outer"-most row and derive the weights of the remaining rows by linear interpolation. These weights are applied to the "accumulated errors" fusion strategy by simply multiplying the distances obtained for each row by the corresponding weight.

# 3 EXPERIMENTAL STUDY

## 3.1 Setting and Methods

For all our experiments we considered images with 8-bit grayscale information per pixel from the CASIA[1] v1.0 iris image database. We applied the experimental calculations on the images of 108 persons in the CASIA database using 7 iris images of each person which have all been cropped to a size of $280 \times 280$ pixels.

The employed iris recognition system builds upon Libor Masek's MATLAB implementation[2] of a 1D version of the Daugman iris recognition algorithm. First, this algorithm segments the eye image into the iris and the remainder of the image ("iris detection"). Subsequently, the iris texture is converted into a polar iris image. Additionally, a noise mask is generated indicating areas in the iris polar image which do originate from eye lids or other non-iris texture noise.

Our MATLAB implementation uses the extracted iris polar image ($360 \times 65$ pixels) for further processing and applies the LTP algorithm to it. Following the

[1] http://www.sinobiometrics.com

[2] http://www.csse.uwa.edu.au/~pk/studentprojects/libor/sourcecode.html

suggestion in (Du et al., 2006), we discard the upper and lower three lines of the LTP polar image due to noise often present in these parts of the data (resulting in a $360 \times 59$ pixels LTP patch). The 1D and 2D signatures described in the last section are then extracted from these patches.

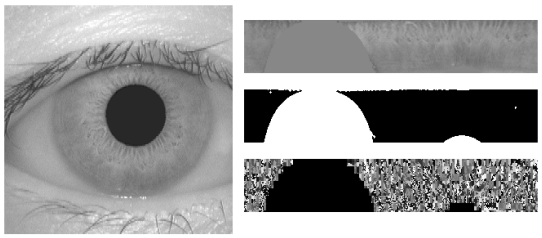


Figure 1: CASIA iris image and the corresponding iris template, noise mask, and LTP patch.

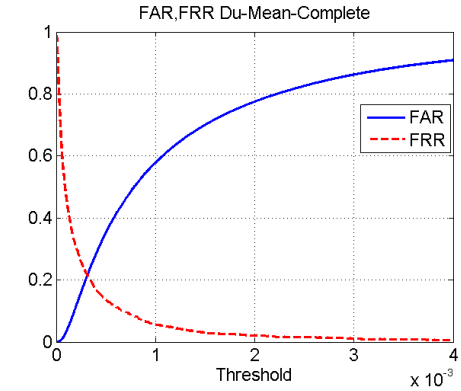
Figure 1 shows an example of an iris image of one person (CASIA database), together with the extracted polar iris image, the noise mask, and the LTP patch (template, noise mask, and LTP patch have been scaled in y-direction by a factor of 4 for proper display).

## 3.2 Experimental Results

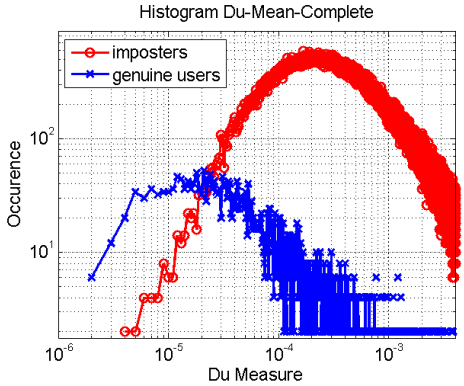
In Figure 2.a, we show the ROC curve of the original version of the Du approach employing 1D signatures based on LTP row mean vectors. EER is rather high with 0.22 and especially the concave shape of the FAR curve for the Du algorithm depicts a steep slope close to zero which means that low FAR values cause unrealistically high FRR. The latter result illustrates the reason why this algorithm is restricted to the similarity ranking scenario in the original work (Du et al., 2006).

The reasons for the respective behaviour can be seen in Figure 2.b. The overlap between genuine users and imposters distributions is is very large for the Du approach, obviously causing the high EER.

When turning to 2D signatures, we compare different fusion strategies and histogram resolutions in Table 1 with respect to their EER. While it is obvious that too many histogram bins lead to poor results (important histogram properties are concealed by noise), also a reduction to 20 bins results in lower EER as compared to 100 bins. When comparing the two fusion strategies, accumulating distances (AD) at a row basis is clearly superior to simple histogram concatenation (HC) at a reasonable histogram resolution. In this scenario, we are clearly able to improve EER as compared to the original Du algorithm (from 0.22 down to 0.16).




(a) ROC-plot: ERR 0.22




(b) Genuine users and imposters distributions

Figure 2: Behaviour of the original DU algorithm.

Table 1: EER for two assessment variants and different histogram resolutions (2D signatures).

| # bins | 1500450 | 255 | 100 | 20 |
|---|---|---|---|---|
| HC | 0.3 | 0.2 | 0.18 | 0.19 |
| AD | 0.32 | 0.16 | 0.16 | 0.18 |

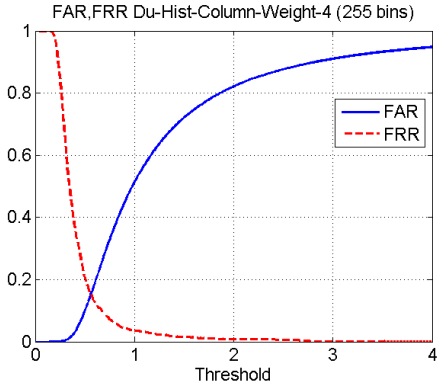
Note also, that histgram resolution up to 255 is beneficial for accumulating errors fusion while it is not for histogram concatenation. This is an intuitive result, since in case of histogram concatenation the vectors to be compared in the Du measure are already fairly long overall, while this is not the case for accumulating errors fusion.

Table 2 compares three weighting strategies for the accumulated errors fusion strategy. The best results are obtained when using weight 4 for the LTP row closest to the pupil. This result confirms the assumption, that "inner" iris information is most important for recognition purposes.

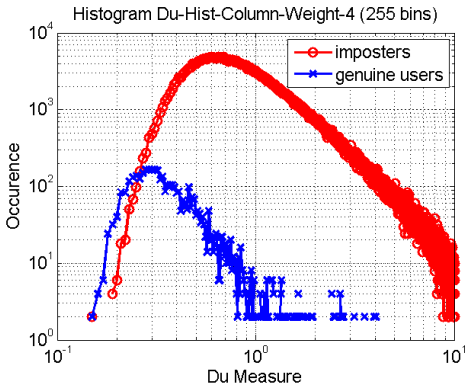
We display the ROC curve for the best setting for accumulted error fusion strategy in Figure 3.a. The graph exhibits a much better behaviour of the FAR

Table 2: EER for three weighting variants and different histogram resolutions (2D signatures).

| histogram bins | 255 | 100 | 20 |
|---|---|---|---|
| no weight | 0.16 | 0.16 | 0.18 |
| weight 2 | 0.15 | 0.15 | 0.19 |
| weight 4 | 0.15 | 0.15 | 0.16 |




(a) ROC-plot: EER 0.15




(b) Genuine users and imposters distributions

Figure 3: Behaviour of best 2D Du variant (accumulated errors (weight 4, 255 bins)).

curve in proximity of zero as compared to the original one which documents also the improved behaviour.

Finally, we visualize genuine users and imposters distributions the same 2D variant of the Du algorithm in Figure 3.b which confirms improvements with respect to the original algorithm.

# 4 CONCLUSIONS

In this work we have improved an iris recognition algorithm based on 1D signatures extracted from the spatial domain by including histogram based information instead of mean values. While we succeeded in maintaining rotation invariance in our improved version, FAR and FRR are still significantly worse compared to state of the art identification techniques which limits this improvement to the employment in a similarity ranking scheme as it is the case for the original version.

One reason for the still disappointing behaviour is as follows: when shifting the different rows in the polar iris image with a different amount against each other, the 2D signatures (as well as the 1D signatures of course) are preserved. This operation corresponds to the rotation of concentric circles of iris pixels by an arbitrary amount – still, the signatures for all those artificially generated images are identical. Our results indicate that indeed information about the spatial position of frequency fluctuations in iris imagery is crucial for effective recognition.

# ACKNOWLEDGEMENTS

Most of the work described in this paper has been done in the scope of a semester project in the master program on “Communication Engineering for IT” at Carithia Tech Institute.

# REFERENCES

Daugman, J. (2004). How iris recognition works. *IEEE Transactions on Circiuts and Systems for Video Technology*, 14(1):21–30.

Du, Y., Ives, R., Etter, D., and Welch, T. (2006). Use of one-dimensional iris signatures to rank iris pattern similarities. *Optical Engineering*, 45(3):037201–1 – 037201–10.

Ives, R., Guidry, A., and Etter, D. (2004). Iris recognition using histogram analysis. In *Conference Record of the 38th Asilomar Conference on Signals, Systems, and Computers*, volume 1, pages 562–566. IEEE Signal Processing Society.

Ko, J.-G., Gil, Y.-H., Yoo, J.-H., and Chung, K.-I. (2007). A novel and efficient feature extraction method for iris recognition. *ETRI Journal*, 29(3):399 – 401.

Ma, L., Tan, T., Wang, Y., and Zhang, D. (2004). Efficient iris recognition by characterizing key local variations. *IEEE Transactions on Image Processing*, 13:739–750.

Zhu, Y., Tan, T., and Wang, Y. (2000). Biometric personal identification based on iris patterns. In *Proceedings of the 15th International Conference on Pattern Recognition (ICPR’00)*, volume 2, pages 2801–2804. IEEE Computer Society.

# PATH PLANNING USING DISCRETIZED EQUILIBRIUM PATHS
## *A Robotics Example*

Cornel Sultan
*Aerospace and Ocean Engineering Department, Virginia Tech, 215 Randolph Hall, Blacksburg, VA, U.S.A.*
*csultan@vt.edu*

Keywords: Nonlinear control, equilibrium path, robot path planning.

Abstract: A collision avoidance path planning problem is considered and a simple solution which uses piecewise constant controls generated by discretizing a feasible equilibrium path is presented and investigated.

## 1 INTRODUCTION

A new methodology has been recently proposed (Sultan, 2007) for the control of nonlinear ODEs,

$$\dot{x} = \frac{dx}{dt} = f(x,u), \quad x \in X \subset R^n, \;\; u \in U \subset R^m, \;\; t \in T \subset R. \tag{1}$$

Here $f$ is a function of class $C^k$ in $X \times U$ ($k > 0$), $x$, $u$, and $t$ are the state, control vectors, and time, whereas $X$, $U$, and $T$ are *open sets* in the $n$, $m$, and one dimensional real spaces.

The key idea is to control (1) such that its *state space trajectory is close to an equilibrium path* obtained by solving

$$0 = f(x,u). \tag{2}$$

If $(x_i, u_i)$ is a solution of (2) and $J_i = \frac{\partial f}{\partial x}(x_i, u_i)$ is *not singular*, there exist an open set $U_e$ and a *unique* function $g$ of class $C^k$ such that

$$x = g(u), x_i = g(u_i), \quad f(g(u),u) = 0, \mathrm{g}: U_e \to X_e. \tag{3}$$

Here $U_e$ is the *largest domain* in $U$ in which (2) can be solved uniquely for $x$ as in (3) and $\frac{\partial f}{\partial x}(g(u),u)$ is not singular. If $(x_f, u_f)$, $u_f \in U_e$ is a different solution of (3), $u_i$ and $u_f$ can be connected by a curve $u_e(s)$ in $U_e$, parameterized by $s \in [0,\tau]$,

$$u_e(0) = u_i, u_e(\tau) = u_f, \tag{4}$$

which is $g$-mapped onto an *equilibrium path,* $x_e(s) = g(u_e(s))$, $x_e(0) = x_i$, $x_e(\tau) = x_f$.

The control problem is to develop control laws which guarantee that *the state space trajectory of the system is close to the equilibrium path*, as illustrated in Figure 1. In order to achieve this goal, the strategy described next was proposed in (Sultan, 2007). The controls are initially fixed at $u_i$ and when the transition begins, at *t=0*, they start to vary along $u_e$, $u(t) = u_e(t)$, $t \in [0,\tau] \subset T$. When $t$ reaches $\tau$ the controls are frozen at the final, desired value:

$$u(t) = \begin{cases} u_i, t < 0 \\ u_e(t), 0 \le t \le \tau \\ u_f, t > \tau \end{cases}. \tag{5}$$

The corresponding state space trajectory, $x_d(t)$, called the *deployment path*, is the solution of

$$\dot{x}_d = f(x_d, u(t)), x_d(0) = x_i. \tag{6}$$

If $x_d(\tau)$ belongs to the basin of attraction of $x_f$ then the system's trajectory will settle down, asymptotically in time, to the desired final value, $x_f$. *Asymptotical stability of* $x_f$ *is crucial* for the application of this methodology.

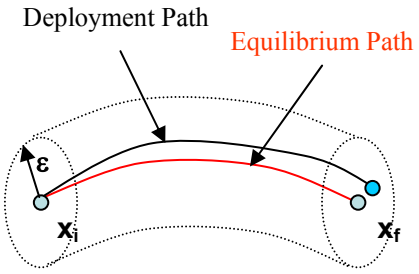



Figure 1: Deployment and equilibrium paths.

In this paper an example of a collision avoidance path planning problem is considered. An equilibrium path which satisfies the constraints is found and discretized to generate *piecewise constant* controls which are used to drive the system. It is important to remark that this strategy is different from the one proposed in (Sultan, 2007), where continuous controls are used. Here, the parameterization of the equilibrium path, originally continuous, is discretized. One justification for this approach is the easiness of discrete controls implementation.

## 2 THEORETICAL RESULTS

In the following two important results are given (the proofs are omitted for brevity).

**Theorem 1.** If the equilibrium path is composed only of *asymptotically stable* equilibria then, for $\forall \varepsilon > 0$ there exists a piecewise constant control $u(t)$, *obtained by discretizing the equilibrium path*, such that the distance between the corresponding segments of the deployment and equilibrium paths is less than $\varepsilon$ (i.e. the deployment and equilibrium paths are arbitrarily close).

**Theorem 2.** If the equilibrium path is composed only of asymptotically stable equilibria and for any $u$, $f(x,u)$ is Taylor series expandable in $x$, for $\forall \eta > 0$ there exists a piecewise constant control $u(t)$, *obtained by discretizing the equilibrium path* such that $\|\dot{x}_d(t)\| < \eta, \forall t \in [0,\tau]$.

## 3 A PATH PLANNING PROBLEM

Consider a two link robotic manipulator in the vertical plane (Figure 2). The links are rigid, the system is placed in a constant gravitational field, control torques and damping torques proportional to the relative angular velocity between the moving parts act at the joints. The equations of motion are:

$$(m_1c_1^2 + m_2l_1^2 + I_1)\ddot{\theta}_1 + m_2l_1c_2\cos(\theta_1-\theta_2)\ddot{\theta}_2 + d_1\dot{\theta}_1 + m_2l_1c_2\sin(\theta_1-\theta_2)\dot{\theta}_2^2 + (m_1c_1+m_2l_1)g\sin(\theta_1) = u_1 \quad (7)$$

$$m_2l_1c_2\cos(\theta_1-\theta_2)\ddot{\theta}_1 + (m_2c_2^2 + I_2)\ddot{\theta}_2 + d_2(\dot{\theta}_2-\dot{\theta}_1) - m_2l_1c_2\sin(\theta_1-\theta_2)\dot{\theta}_1^2 + m_2c_2g\sin(\theta_2) = u_2 \quad (8)$$

where angles $\theta_1, \theta_2$ describe the motion, $m_i$, $l_i$, $c_i$, $I_i$, are the mass, length, center of mass (CM) position, transversal moment of inertia of the $i$-th link, $d_i$ and $u_i$ are the damping coefficient and control torque at joint $i$, respectively, $g$ is the gravitational constant. These equations can be easily cast into the first order form (1). The numerical values (SI units) used are:

$$m_1 = 10, m_2 = 5, l_1 = l_2 = 1/\sqrt{3}, c_1 = c_2 = 0.5, I_i = m_i l_i / 12, d_1 = d_2 = 0.5, g = 9.81. \quad (9)$$

The system must transition between two equilibria, $\theta_{1i} = 70, \theta_{2i} = 0, \theta_{1f} = -70, \theta_{2f} = 0$.

Collision with a circular sector obstacle, of radius $R=1$, described below, must be avoided:

$$\begin{aligned} &\theta_2 > 60 + \theta_1, if \quad -60 < \theta_1 \le 0 \\ &\frac{\sin(30-\theta_2)}{\sin(\theta_1-\theta_2)} - \frac{1}{\sqrt{3}} < 0, if \quad 0 < \theta_1 < 30 \\ &\theta_2 > 60 - \theta_1, if \quad 30 \le \theta_1 < 60. \end{aligned} \quad (10)$$

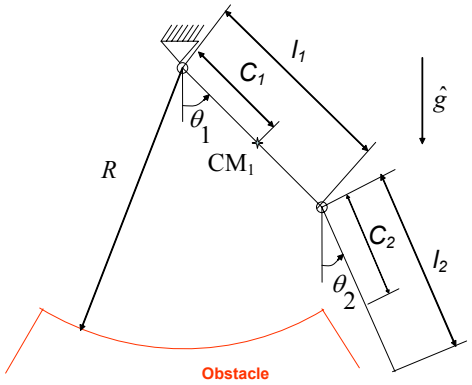



Figure 2: Two link robotic manipulator.

An equilibrium path which satisfies (10) is

$$\theta_{2e} = 62\cos\left(\frac{\theta_{1e}}{\theta_{1f}-\theta_{1i}}\right), \theta_{1e} \in [\theta_{1i}, \theta_{1f}]. \quad (11)$$

and the equilibrium controls are easily found,

$$\begin{aligned} u_{1e} &= g(m_1 c_1 + m_2 l_1)\sin(\theta_{1e}), \\ u_{2e} &= m_2 c_2 g \sin(\theta_{2e}). \end{aligned} \quad (12)$$

The equilibrium path is parameterized using the following class $C^2$ function

$$\begin{aligned} \theta_{1e}(t) &= \theta_{1i} + \frac{30}{\tau^5}(\theta_{1f}-\theta_{1i}) \\ &\left(\frac{t^5}{30} - \frac{t^4}{6}(t-\tau) + \frac{t^3}{3}(t-\tau)^2\right), 0 \le t \le \tau, \end{aligned} \quad (13)$$

which is further discretized to obtain piecewise constant controls using (11) and (12).

Consider $\tau = 10$ ("fast deployment"). Piecewise constant controls are generated using N equal time intervals. Figure 3 shows the deployment and equilibrium paths.

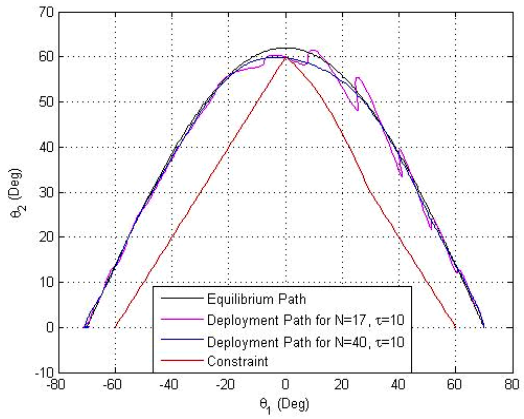



Figure 3: Deployment paths for "fast" deployment.

Figures 4 and 5 give the time histories of the controls and angles for N=17 and N=40. *The deployment error cannot be made small enough to avoid the obstacle* regardless of how large N is (higher values of N were considered). Thus $\tau$ *should be increased and the controls refined* for the deployment error to be sufficiently small.

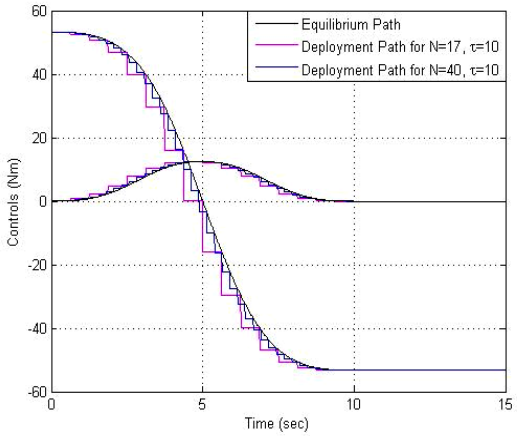



Figure 4: Controls variation for "fast" deployment.

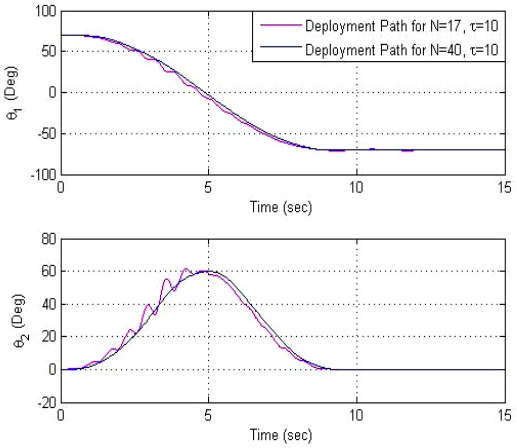



Figure 5: Generalized coordinates variation for "fast" deployment.

In the second scenario, called "slow" deployment, the deployment time is $\tau = 20$ and piecewise constant controls are generated by discretizing (11-13) with N=34 and N=80. Figures 8-10 show that collision is avoided. The deployment error is smaller because the deployment time is longer and finer controls are used. It is important to mention that if only the deployment time is increased the desired result is not obtained; if N=17 or N=40 are used in conjunction with $\tau = 20$, the deployment error is still big and collision with the obstacle occurs.

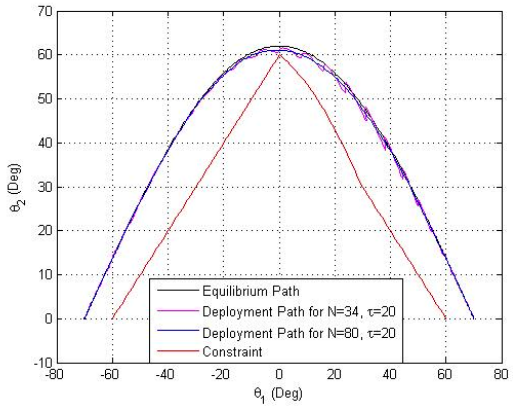



Figure 6: Deployment paths for "slow" deployment.

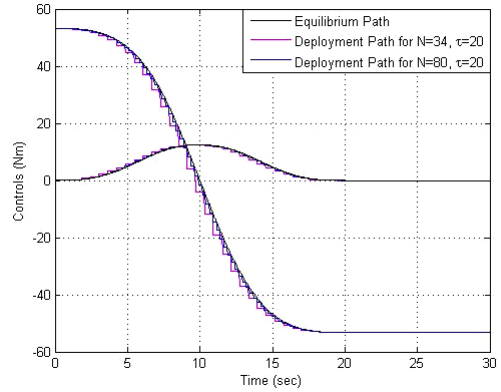



Figure 7: Controls variation for "slow" deployment.

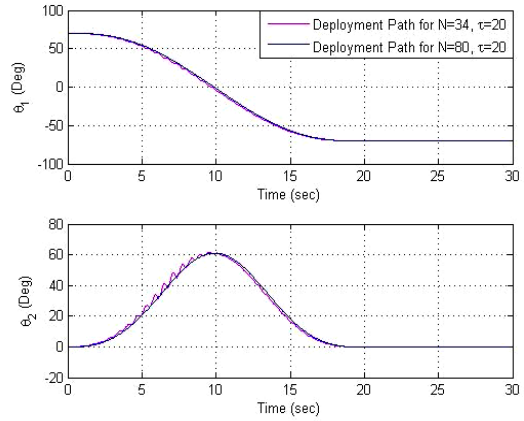



Figure 8: Generalized coordinates variation for "slow" deployment.

# 4 CONCLUSIONS

An example of a path planning problem is used to illustrate the control of nonlinear systems using equilibrium paths. The idea is to find an equilibrium path which satisfies the collision avoidance constraints, which is a much easier problem than finding a dynamic path which satisfies the constraints. Then the equilibrium path is discretized to build piecewise constant controls which are used to drive the system. Simulations indicate that for the deployment and equilibrium paths to be close the deployment time should be sufficiently long and the controls sufficiently refined.

It is important to remark that the solution investigated here uses discretizations of an equilibrium path which satisfies the collision avoidance constraints as opposed to continuous parameterizations and hence continuous controls. One justification for this approach is the easiness of practical implementation of discrete controls.

# REFERENCES

Sultan, C., 2007. Nonlinear systems control using equilibrium paths. In *Proceedings of the Conference on Decision and Control, New Orleans, LA, USA.*

# A FRAMEWORK FOR DISTRIBUTED AND INTELLIGENT PROCESS CONTROL

Qurban A. Memon
*UAE University, United Arab Emirates*
*qurban.memon@uaeu.ac.ae*

Keywords: Distributed network control, Process Control, Intelligent agents.

Abstract: The customized development of the Distributed Control System for process control in an environment of intelligent and tagged field devices etc., is the main focus of this work. The proposed solution consists of two-layer approach: use of decentralized intelligent agents at the local process level, and four-tier modular architecture at central controller level to help implement distributed intelligence. The design and development issues for such a customized design are investigated.

## 1 INTRODUCTION

The decentralization of control, and expanding physical setups have resulted into today's distributed process control (DCS) systems (M. Ioannides, 2004, J. Alonso, 2000). The research in this area is quite active because of these developments, for example, Profibus fieldbus networks and wireless Profibus for real time industrial control systems (E. Tovar, 1999, A. Willing, 2003). Recently focus is reported with respect to distributed intelligence for reduced operational changes. These efforts have respective generated agent based approach (Bernan, 2002, F. Maturna, 2005). Currently, active Radio Frequency Identification (RFID) devices are being deployed for a variety of process control industry solutions (A. Juels, 2003, J. Bohn, 2004). To some industries, RFID is bringing a level of automation and control similar to what process control devices brought to manufacturing decades ago. I. Satoh, 2004, presented a framework which exploits agents to enhance capabilities of the users in an environment of tagged devices. In another work (S. Naby, 2006), the author discusses idea of integrating software agents into RFID architectures to accumulate information from tags and process them for customer/object or system specific use, for example a concurrent mission. As a summary, the optimization in network performance combined with distributed intelligence in an environment of non-stationary and reconfigurable devices provides a new direction of research.

The process under investigation is shown in Figure 1, which shows reprogrammable and reconfigurable control devices including some tagged and distributed field intelligent devices. In short, the challenges for DCS development include process reconfigurability and intelligent decision making within a Profibus/Profinet compatible network. For comparison, a model is to be used to set a baseline for performance. As the network is distributed, hence a multiple input multiple output (MIMO) baseline is considered that requires performance matching to that of the centralized MIMO. For Figure 1 to achieve similar performance to that of the centralized MIMO, each parameter update needs to be communicated over the network at times. In order to categorize time delay $d(t)$ in the network, we divide it into three categories:

$$d(t) = d_1(t) + d_2(t) + d_3(t) \quad (1)$$

where $d_1(t)$, $d_2(t)$, and $d_3(t)$ represent time delay when a device communicates with controller, time delay when a group of devices communicate with controller; and when all devices communicate with controller respectively. One obvious approach could be to minimize either of these delays so as to optimize the performance to match a centralized MIMO system.

## 2 PROPOSED APPROACH

A set of processes is proposed to introduce intelligence at field level to gain respective

independence. This way, minimized communication with the main controller thwarts communication bottlenecks caused by interoperability of devices, or simple operational requirements at local level. This also improves survivability of the local processes in cases when central controller fails in providing critical timely decision. The configurability of devices may be provided by collecting operational parameters at the device(s) level followed by estimation of parameters of concerned entities at the central level. This leads to two separate domains:

## 2.1 Local Process

The local entities tend to be distributed throughout the environment to support overall operations. The job of these entities can be done effectively by agents. The agents collaborate, learn and adjust their abilities within the constraints of the global process.
Agent design mechanism: A lot of work is done on agents alone and details can be found in (R. Brennan, 2001) Agents are active software entities that can request for additional capabilities once they discover that the task at hand can not be fulfilled. The programming of these agents is done at the central level where a set of heuristics is used for reasoning at the local level, and is stored as a function block diagram (like an internal script). The agents know about their equipment, continuously monitor its state, and can decide whether to participate in a mission or not. The collaborating agents join (on their own will) and thus form a cluster in order to enable a decision making. In addition to agents, there are other computing units that exist at the local level and help to form a cluster. These are known as cluster directory (CD) and cluster facilitator (CF) respectively. The following steps describe agent collaboration in clusters:

- Agent N receives a request from central process.
- It checks its scripts, and solves local steps.
- For external steps, it receives contact details of other agents with external capability from CD.
- Agent N creates $CF_N$ and passes on these details.
- $CF_N$ passes the request to specified agents, and thus cluster is formed.

For efficient collaboration, CD must remain updated for recording the information of its members, such as agent name, agent locator, service name, service type, and so on. Upon joining or leaving the cluster, an agent must register or cancel registration respectively through CF. Through a query, an agent can find out other members' services and locators. Through these steps, a trust is developed and thus members hold higher authority than non-members.
Recently developed tools may be used to help design cluster facilitator (CF) and domain ontology, using for example DARPA Markup Language (DARPA, DAML, 2000). The DAML extends XML (Extensible Markup Language) and RDF (Resource Description Framework) to include domain ontology. It provides rich set of constructs to create ontology and to markup information for attaining machine readability and understandability. Furthermore, the Foundation for Intelligent Physical Agent (FIPA, 2003) Agent Management Specification is extended to develop the agent role called CF to manage cluster directory (CD) and cluster ontology. Using assistance from DAML-based ontology, the members of the cluster are able to form cluster and communicate with other agents.

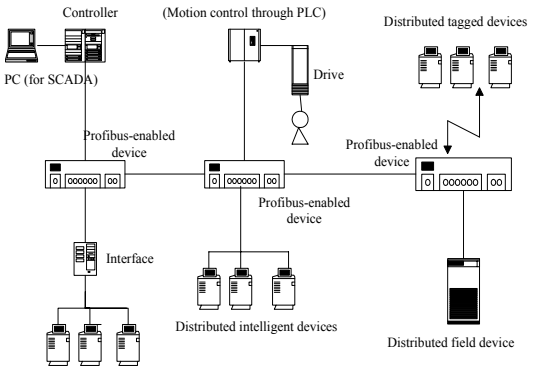



Figure 1: Typical Process Control Network.

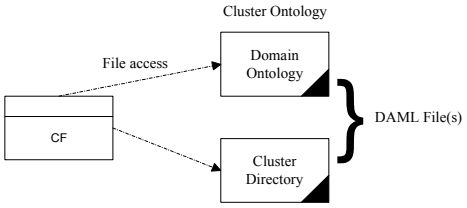



Figure 2: Linking CF with DAML.

The interaction among domain ontology, CD and CF can be best understood using Figure 2. Figure 2 shows how CF gets access to DAML files and facilitates the common goal of the cluster. There are tools available like Jena semantic web that can be used to handle the cluster director (CD) built using DAML, and to develop a Java class "Directory". Thus, main functions of CD can be summarized, as:

- Add and Remove the information of an agent
- Get the list of agent names of all members
- Get the information of individual agent by name

- Get ontology used by members in the cluster
- Add external ontology if provided by an agent

Using local process mechanism and main functions of CD, the partial directory can be described as shown in Figure 3. It shows information of CF (lines 1-9) and members of cluster (lines 20-22), the cluster directory also records meta-data about cluster such as cluster name (line 12), cluster description (lines 13-15), ontology used in cluster (lines 16-18), etc.

An example can be illustrated to show how ontology may be updated (Fig. 4(b)) and that how interactions may develop in a local process. It should be noted here that basic cluster ontology provided by CF remains the same but all members' domain knowledge (ontology) may not be the same. For example, user agent holds basic knowledge of the local process but does not understand the knowledge that a distributed field device holds. Through DAML-based ontology, members can communicate with each other to acquire requested service, as shown in Figure 4. It is clear from Figure 4 that when distributed field device agent joins the cluster, it informs CF about corresponding ontology it provides (Figure 4(a)). Thus the CF maintains local process ontology plus the distributed field device ontology. When a user agent wants to perform a task, it asks CF about domain ontology and the agents that provide external capability. In response, CF informs the user agent if ontology is to be acquired (Figure 4(c)). Thus, the user agent can communicate with the distributed field device agent (Figure 4(d)).

## 2.2 Central Process

This process embodies core, like definition of controller tasks, and definition of domain ontology of each cluster. The other components are removal of agent deadlocks, estimation of local characteristics and decision making in cases when situation develops beyond the capabilities of agent clusters. It can be argued that if only small scale changes are to be decided at the central level like reconfiguration of device processes then intelligence can further be distributed to the agents at local level. In Figure 5, the model architecture of four tiers is shown to implement objectives of the central system. At the bottom layer (Tier 1), active readers or Profibus/Profinet enabled devices collect data, often collected on a trigger similar to a motion sensor. These readers should be controlled by one and only one edge server to avoid problems related to network partitioning. This layer also provides hardware abstraction for various Profibus/Profinet compatible hardware and network drivers for interoperability of devices. The edge sever (Tier 2) regularly poll the readers for any update from device agents, monitors tagged devices and distributed devices through readers, performs device management, and updates integration layer. This layer may also work with system through controls and open source frameworks that provide abstraction and design layer. The integration layer (Tier 3) provides design and engineering of various objects needed for central controller as well as for field processes and for simulation levels of reconfigurability. This layer is close to business application layer (Tier 4). The monitoring of agents behavior, its parameters and cluster characteristics are done at this layer to assess the degree of reconfigurability.

```
1. <cluster:CF rdf:ID="theCF">
2. <cluster:agentName>"CF"</cluster:agentName>
3. <cluster:agentDescription>
4. "DCS Cluster Facilitator"
5. </cluster:agentDescription>
6. <cluster:locator>
7. "http://dcs.ee.uaeu.ac.ae/DCS/agent/CF"
8. </cluster:locator>
9. </cluster:CF>
10.
11. <cluster:Cluster rdf:ID="DCSCluster">
12.<cluster:clusterName>"DCS"</cluster:clusterName>
13. <cluster:clusterDescription>
14. "Distributed Control System"
15. </cluster:clusterDescription>
16. <cluster ontology>
17. "http://dcs.ee.uaeu.ac.ae/DCS/ontology/dcs.daml"
18. </cluster:ontology>
19.
20. <cluster:hasCF rdf:Resource="#theCF"/>
21. <cluster:consistOf rdf:Resource="#agent1"/>
22. <cluster:consistOf rdf:Resource="#agent2"/>
23. </cluster:Cluster>
```

Figure 3: DCS Cluster Directory.

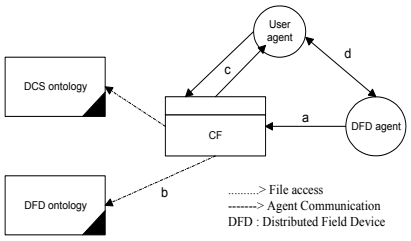



Figure 4: Ontology update provided by DFD.

This layer also takes care of parameters like handling device processes, resource allocation and scheduling of processes. The separation of edge server and integration layer improves scalability and reduces cost for operational management, as the edge is lighter and less expensive. The processing at the edge reduces data traffic to central point. Similarly, the separation of integration from business applications helps in abstraction of process entities.

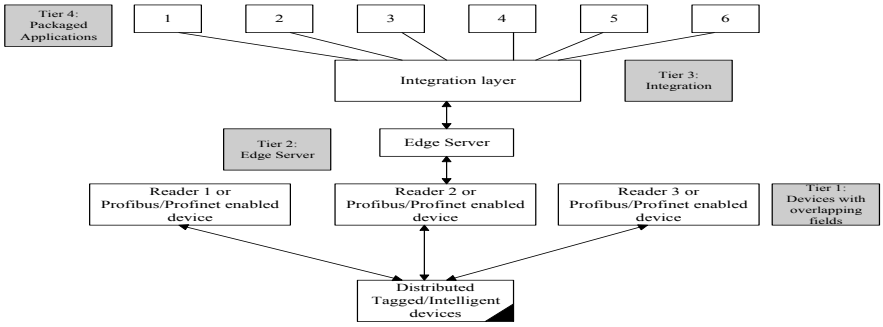



Figure 5: 4-Tier Reference Architecture.

The Tier 3 also enables it as self-healing and self-provisioning service architecture to increase availability and reduce support cost. Control messages flow into the system through business application portal to the integration layer, then on to the edge and, eventually, to the reader. Provisioning and configuration is done down this chain, while reader data is filtered and propagated up the chain.

# 3 CONCLUSIONS

The main idea behind two processes is decentralization. The communication delay is reduced at the cost of increased intelligence at the local level. In fact, if we look at equation (1) we see that $d_1(t)$, $d_2(t)$ and $d_3(t)$ minimize to a level when problem of the node device exceeds the threshold level of the agent intelligence. If collaborative intelligence exceeds combinatorial complexity then there is no need of communication between devices and the controller and requirements of the central process reduce to that of the design of agents only. Thus, the performance matches to that of the centralized MIMO system. The four-tier modular architecture at central level helps in implementation of distributed intelligence at field level and in designing of agents. The functionality more appropriate to the layer has been fit into respective tiers at central level. Additionally, design and reconfigurability can help introduce features in agents to thwart intrusive agents, during real time. This set of gains has not been claimed in either of the approaches (E. Tovar, 1999, A. Willing, 2003).

# ACKNOWLEDGEMENTS

This work was financially supported by UAE University under a grant no. 01-04-7-11/07.

# REFERENCES

A. Juels, R. Rivest, M. Szydlo, “The Blocker Tag: Selective Blocking of RFID Tags for Consumer Privacy”, *Proceedings of ACM conference on Computer and Communications Security*, 2003.

A. Willig, “Polling based MAC protocols for improving real-time performance in a wireless Profibus”, *IEEE Trans. on Industrial Electronics*, pp. 806-817, 2003.

Bernnan, M. Fletcher, and D. H. Norrie, “An agent-based approach to reconfiguration of the real-time distributed control systems,” *IEEE Transactions on Robotics and Automation*, Vol. 18, No. 4, 2002.

DARPA Agent Markup Language, *http://www.daml.org/*

E. Tovar, and F. Francisco, “Real-time field bus comm. using Profibus Networks”, *IEEE Trans. on Industrial Electronics*, pp. 1241-1251, 1999.

F. Maturana, R. Staron, K. Hall, “Methodologies and Tools for Intelligent Agents in Distributed Control”, *IEEE Intelligent Systems*, pp. 42-49, February 2005.

Foundation for Intelligent Physical Agents (FIPA), *http://www.fipa.org/specs/fipa00023/*

I. Satoh, “Software agents for ambient intelligence”, *Proceedings of IEEE International Conference on Systems, Man and Cybernetics*, pp.1147-1150, 2004.

J. Alonso, et al, “Development of a Distributive Control Scheme for Fluorescent Lighting based on LonWorks Technology”, *IEEE Transactions on Industrial Electronic*, Vol. 47, No. 6, pp. 1253-1262, 2000.

J. Bohn, F. Mattern, “Super-Distributed RFID Infrastructures”, *Lecture Notes in Computer Science (LNCS)* No. 3295, Springer-Verlag, pp. 1-12, Eindhoven, Netherlands, November 8-10, 2004.

M. G. Ioannides, “Design and Implementation of PLC based Monitoring Control System for Induction Motor”, *IEEE Transactions on Energy Conversion*, pp. 469-476, 2004.

R. Brennan and D. Norrie, “Agents, holons and function blocks: Distributed intelligent control in manufacturing”, *Journal of App. Sys Studies*, 2001.

S. Naby, and P. Giorgini, “Locating Agents in RFID Architectures”, Technical Report # DIT-06-095, University of Trento, Italy, December 2006.

# HYBRID WAVELET-KALMAN FILTER MULTI-SCALE SEQUENTIAL FUSION METHOD

Funa Zhou[1,2] and Tianhao Tang[1]
[1.] *Department of Electrical & Control Engineering, Shanghai Marintime University, Shanghai, China*
[2.] *Computer&Information Engineering School, Henan University, Kaifeng, Henan, China*
*Zhoufn2002@163.com, thtang@cen.shmtu.edu.cn*

Keywords: Hybrid wavelet-Kalman filter, Sequential fusion, Non-$2^n$ sampling.

Abstract: With the development of automation, multi-scale data fusion has become a hot research topic, however, limited by the constraint that signal to implement wavelet transform must have the length of $2^q$, multi-scale data fusion problem involved with non-$2^n$ sampled observation data still hasn't been efficiently solved. In this paper, we develop a hybrid wavelet-Kalman filter multiscale sequential fusion method. First, we develop the hybrid wavelet-Kalman filter multiscale estimation method which combines the advantage of wavelet and Kalman filter to obtain the real time, recursive, multiscale estimation of the dynamic system. Then, a multiscale sequential fusion method is presented. Under the hybrid wavelet-Kalman filter multiscale estimation frame, we can easily fuse information from multiple sensors sequentially without designing other complex fusion algorithm. The multiscale sequential fusion method can fuse non-$2^n$ sampled data just by analyzing the possible observation structure to design the observation model of the stacked dynamic system. Simulation result of three sensors with sampling interval 1, 2 and 3 shows the efficiency of this method.

## 1 INTRODUCTION

In many fields, such as, automatic control, aerospace, communication, navigation and production industry, more than one sensor is used to gather complete information of the object or process. According to the mechanism of each sensor, they can be placed on different scales and the sampling rate of these sensors may also be different. The research of multi-sensor data fusion for dynamic system is significant both in practice and theoretically (Wen 2002a, Wen 2002b, Lang Hong1994). Especially, in many cases, the sampling interval may not equal to $2^n$, thus it is inconvenient for us to fuse information provided by these sensors. Therefore the tracking or estimation accuracy may be strongly reduced.

The main technique used in multi-scale data fusion is Kalman filter and wavelet analysis. Kalman filter can result in real-time, recursive and optimal estimate while it doesn't take the multi-scale character of the object into account. Wavelet transform can implement multi-scale analysis and estimation of the dynamic system, but the estimate is neither real time nor recursive (Wen 2002a).

Using Kalman filter, data fusion algorithm for multi-sensor sampling at same rate has been successfully developed. Coporating with multi-scale theory, multi-scale data fusion algorithm for multi-sensor sampling at $2^n$ interval has also been developed. Limited by the fact that signals to implement wavelet transform must have the length of $2^q$, the method mentioned in (Wen 2002b, Lang Hong1994) can't solve the data fusion problem when the sensors used are not sampling at $2^n$ interval.

We find that once the dynamic system is stacked in a given length $2^q$, sensors not sampled at interval $2^n$ has different observation structure on each block, that is, the length of the observation vector on each block may be different, and the sampling rule on each observation block is also different.

Based on the hybrid wavelet and Kalman filter sequential fusion method developed in (Wen 2006a, Zhou 2007), we are intend to develop a sequential fusion scheme by designing the stacked observation model to fuse the observation data coming from those sensors sampling at non-$2^n$ interval.

# 2 MANUSCRIPT PREPARATION

## 2.1 Dynamic System

Considering a system involving $K$ sensors

$$x(k+1) = A(k)x(k) + w(k) \quad (1)$$

$$z_i(k_i) = C_i(k_i)x(k_i) + v_i(k_i) \ i = 1,2,\cdots,K \quad (2)$$

where $k \in N$, $k_i = d_i k$, $d_i \in N$ is the sampling interval of each sensor, $x(k) \in R^n$ is the object's state, $A(k) \in R^{n\times n}$ is the system matrix.

The System's process noise $w(k) \in R^n$ is the Gaussian white noise with the following statistics

$$E\{w(k)\} = 0 \quad (3)$$

$$E\{w(k)w^T(l)\} = Q(k)\delta_{k,l} \quad k,l \geq 0 \quad (4)$$

$Q(k)$ is a nonnegative matrix.

The observation noise $v_i(k_i)$ is also Gaussian white noise

$$E\{v_i(k_i)\} = 0 \quad (5)$$

$$E\{v_i(k_i)v_j^T(l_j)\} = Q(k)\delta_{i,j}\delta_{k,l} \quad i,j = 1,2,\cdots,K \quad (6)$$

$x(0)$ is the initial state of the system,

$$E\{x(0)\} = x_0 \quad (7)$$

$$E\{[x(0)-x_0][x(0)-x_0]^T\} = P_0 \quad (8)$$

$x(0)$, $w(k)$ and $v_i(k_i)$ is independent.

## 2.2 Stacked System

Rewrite the dynamic model (1) and (2) as a stacked system with block length $M$

$$\overline{X}(m+1) = \overline{A}(m)\overline{X}(m) + \overline{W}(m) \quad (9)$$

$$\overline{Z}_i(m) = \overline{C}_i(m)\overline{X}(m) + \overline{V}_i(m) \ i = 1,2,\cdots K \quad (10)$$

where

$$\overline{X}(m) = [x^T((m-1)M+1),\cdots,x^T((m-1)M+M)]^T \quad (11)$$

$$\overline{A}(m) = \mathrm{diag}[\prod_{j=0}^{M-1}A(mM-j),\cdots,\prod_{j=0}^{M-1}A(mM+M-j)] \quad (12)$$

$$\overline{Z}_i(m) = [z_i^T(M(m-1)+d_i-r_i(m)), z_i^T(M(m-1)+2d_i-r_i(m)), \cdots; z_i^T(M(m-1)+S_i(m)d_i-r_i(m))]^T \quad (13)$$

where $\overline{Z}_i(m)$ is the observation of $m$ th block observed by sensor $i$. $\overline{C}_i(m)$ in equa.(14) is the observation matrix, $r_i(m) = m(M-1) \bmod d_i$, $e(a)$ is the unit vector whose $a$ th element is 1, while other elements are all 0.

$$\overline{C}_i(m) = \begin{bmatrix} C_i(M(m-1)+d_i-r_i(m))\cdot e(d_i-r_i(m)) \\ C_i(M(m-1)+2d_i-r_i(m))\cdot e(2d_i-r_i(m)) \\ \vdots \\ C_i(M(m-1)+S_i(m)d_i-r_i(m))\cdot e(S_i(m)-r_i(m)) \end{bmatrix} \quad (14)$$

Section 4.2 shows the detailed designing of $\overline{C}_i(m)$.

$\overline{V}_i(m)$ is the observation noise with statistics

$$E\{\overline{V}_i(m,s)\} = 0 \quad (15)$$

$$E\{\overline{V}_i(m,s)\overline{V}_j^T(m,t)\} = R_i\delta_{i,j}\delta_{s,t} \ s,t = 1,2,\cdots,S_i(m) \quad (16)$$

where $S_i(m)$ is the length of the observation vector on the $m$ th block. $\overline{W}(m)$ is the process noise

$$\overline{W}(m) = B(m)\widetilde{W}(m) \quad (17)$$

$$\widetilde{W}(m) = [w^T((m-1)M+1),\cdots,w^T((m-1)M+2M-1)]^T \quad (18)$$

$$E[\overline{W}(m)] = 0 \quad (19)$$

$$\overline{Q}(m) \equiv E[\overline{W}(m)\overline{W}^T(m)] = B(m)\widetilde{Q}(m)B^T(m) \neq 0 \quad (20)$$

$$B(m) = \begin{bmatrix} \prod_{j=0}^{M-2}A(mM-j) & \prod_{j=0}^{M-3}A(mM-j) & \cdots & I & 0 & \cdots & 0 \\ 0 & \prod_{j=0}^{M-2}A(mM+1-j) & \prod_{j=0}^{M-2}A(mM+1-j) & \cdots & I & 0 & 0 \\ \cdots & \cdots & \cdots & & \cdots & & \\ 0 & 0 & \cdots & 0 & \prod_{j=0}^{M-2}A(mM+M-1-j) & \cdots & I \end{bmatrix} \quad (21)$$

$$\widetilde{Q}(m) = diag[Q((m-1)M+1),\cdots,Q((m-1)M+2M-1)]^T \quad (22)$$

in equations (14)-(22), $diag[X,Y,\cdots,Z]$ is the blocked diagonal matrix.

## 2.3 Multiscale Stacked System

Implementing wavelet transform on equation (9)

$$W_X\overline{X}(m+1) = W_X\overline{A}(m)\overline{X}(m) + W_X\overline{W}(m) \quad (23)$$

That is

$$\overline{\gamma}(m+1) = \overline{A}_w(m)\overline{\gamma}(m) + \overline{\mu}(m) \quad (24)$$

where $W_X$ is the operator matrix of wavelet transform, satisfying[Wen 2002 a, Lang Hong1994]

$$W_X^* W_X = I$$

$$\overline{\gamma}(m) = W_X \overline{X}(m) \quad (25)$$

$$\overline{\mu}(m) = W_X \overline{W}(m) \quad (26)$$

$$\overline{A}_w(m) = W_X \overline{A}(m) W_X^* \quad (27)$$

$$\overline{Q}_w(m) = W_X \overline{Q}(m) W_X^* \quad (28)$$

It is easy to prove that the process noise of the new stacked system (24) is statistically independent, that is $\overline{Q}_w(m) = 0$ , which is also one of the advantages of hybrid wavelet-Kalman filter: de-correlating the correlation between blocks (Wen 2006a).

With the orthogonality of $W_X$ , we can rewrite the observation equation as

$$\overline{Z}_i(m) = \overline{C}_i(m) W_X^* W_X \overline{X}(m) + \overline{V}_i(m) \quad (29)$$

$$\overline{Z}_i(m) = \overline{C}_i(m) W_X^* \overline{\gamma}(m) + \overline{V}_i(m) \quad i = 1,2,\cdots,K \quad (30)$$

$$\overline{\gamma}(m) = W_X \overline{X}(m) \quad (31)$$

That is

$$\overline{Z}_i(m) = \overline{C}_i(m) W_X^* \overline{\gamma}(m) + \overline{V}_i(m) \quad i = 1,2,\cdots,K \quad (32)$$

# 3 HYBRID WAVELET-KALMAN FILTER MULTI-SCALE ESTIMATION FOR A SINGLE SENSOR

The following state transition equation and the observation equation of the wavelet transform coefficient of the $m$ th block can be established(Tong 2000)

$$\overline{\gamma}(m, s+1) = \overline{\gamma}(m, s) + \overline{w}(m, s), s = 0,1,2,\cdots,S_1 - 1 \quad (33)$$

$$\overline{Z}_1(m, s) = H(m, s)\overline{\gamma}(m, s) + \overline{V}_1(m, s), \quad s = 1,2,\cdots,S_1 \quad (34)$$

where $\overline{Z}(m, s)$ is the observation at time s of block $m$ . In equa.(34),

$$H(m) \equiv \overline{C}(m) W^T \quad (35)$$

where $H(m, s)$ is the $s$ th row of the matrix $H(m)$ .

The main idea of hybrid wavelet-Kalman filter method includes two steps (Wen 2006 a, Wen 2006 b, Zhou 2007):

(1) Wavelet transform coefficients prediction based on stacked dynamic system $\hat{\overline{\gamma}}(m \mid m-1) = \overline{A}_w(m)\hat{\overline{\gamma}}(m-1 \mid m-1)$ .

(2) Use each observation on this block to update the estimation of wavelet transform coefficient.

Implement Kalman filter on the system given by equa. (33) and (34). In each block, the original state can be derived by a prediction process

$$\hat{\overline{\gamma}}_{0|0}(m) = \overline{A}_w(m)\hat{\overline{\gamma}}(m-1 \mid m-1) \quad (36)$$

$$\overline{P}_{0|0}(m) \equiv E[\hat{\overline{\gamma}}_{0|0}(m)\hat{\overline{\gamma}}_{0|0}^T(m)] = \overline{A}_w(m)\overline{P}_w(m-1 \mid m-1)\overline{A}_w^T(m) + \overline{Q}_w(m) \quad (37)$$

The filter process follows as

$$\hat{\overline{\gamma}}_{s+1|s+1}(m) = \hat{\overline{\gamma}}_{s+1|s}(m) + K(s+1)\widetilde{Z}(m, s+1)$$

$$s = 0,1,2,\cdots,S_1 - 1 \quad (38)$$

$$\hat{\overline{\gamma}}_{s+1|s}(m) = \hat{\overline{\gamma}}_{s|s}(m) \quad (39)$$

$$P_{s+1|s}(m) = P_{s|s}(m) \quad (40)$$

$$K(s+1) = P_{s+1|s}(m)H^T(m, s+1)[H(m, s+1)P_{s+1|s}(m) H^T(m, s+1) + \overline{R}(m, s+1)]^{-1} \quad (41)$$

$$\widetilde{Z}(m, s+1) = \overline{Z}(m, s+1) - H(m, s+1)\hat{\gamma}_{s+1|s}(m) \quad (42)$$

$$P_{s+1|s+1}(m) = [I - K(s+1)H(m, s+1)]P_{s+1|s}(m) \quad (43)$$

This filter process is essentially the gradually updating of the prediction $\hat{\overline{\gamma}}_{0|0}(m)$ .The final updating as the estimation of this block

$$\hat{\overline{\gamma}}(m \mid m) = \hat{\overline{\gamma}}_{S_1|S_1}(m) \quad (44)$$

$$\overline{P}_w(m \mid m) = \overline{P}_{S_1|S_1}(m) \quad (45)$$

The whole process of hybrid wavelet-Kalman filter method can be shown in figure 1.

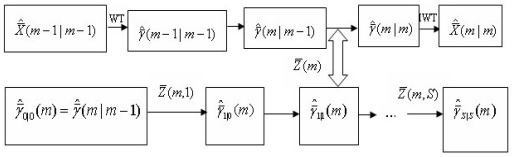



Figure 1: Hybrid wavelet-Kaman filter Algorithm.

# 4 NON-$2^n$ SAMPLED DATA'S SEQUENTIAL FUSION

## 4.1 Sequential Fusion based on Hybrid Wavelet-Kalman Filter

To fuse the observation data coming from multiple sensors we can simply cascade these data sequentially. Then use the cascaded data to update

the prediction $\hat{\bar{\gamma}}_{0|0}(m)$ more times than only one sensor case. The total updating times is

$$S = \sum_{i=1}^{K} S_i \tag{46}$$

This $S$ times updating is the fusion estimation of the wavelet transform coefficient. The sequential fusion process can be shown in fig.2.

The main advantage of this sequential fusion is that the fusion estimate process uses the same mechanism with that of the hybrid wavelet-Kalman filter in one single sensor case without designing a new complex fusion rule.

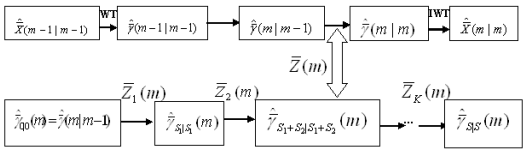



Figure 2: Hybrid wavelet-Kalman filter Sequential fusion.

This sequential fusion algorithm doesn't require that the sampling interval of the observation is $2^n$ .Thus we can manage to process the fusion problem involving non-$2^n$ sampling data

## 4.2 Blocked Observation Model for Non-$2^n$ Sampling Case

One crucial step in hybrid wavelet-Kalman filter is to determine the structure of the stacked observation matrix $\bar{C}_i(m)$ especially for the non-$2^n$ sampling case since the observation structure and observation vector of each block are different.

By analyzing, we find that $\bar{C}_i(m)$ varies periodically with $m$ .The varying rule is determined by the sampling interval $d_i$ and the block length $M$ . The varying period is the minimum common multiple of $M$ and $d_i$ .

For clarity, we display the observation structure in the case $M = 4$ and $d_i = 3$ , $M = 8$ and $d_i = 3$ for the system $n = 1$, $A(k) = A$, $C_i(k_i) = C_i$.

For $M = 4$ and $d_i = 3$ , the period of $\bar{C}_i(m)$ is 12, that is 3 blocks. In these 3 blocks, sensor $i$ samples 4 observation data in total.

$$\bar{C}_i(m) = \begin{cases} C_i \cdot [0\ 0\ 1\ 0] & m = 3j-2 \\ C_i \cdot [0\ 1\ 0\ 0] & m = 3j-1 \\ C_i \cdot [1\ 0\ 0\ 0; 0\ 0\ 0\ 1] & m = 3j \end{cases} \tag{47}$$

where semicolon denotes another row in the matrix. Equation (59) means that in the $m = 3j-2$ block, sensor $i$ sampled 1 observation data; in the $m = 3j-1$ block, sensor $i$ sampled 1 observation data; in the $m = 3j$ block, sensor $i$ sampled 2 observation data.

For the case $M = 8$ and $d_i = 3$ , the period of $\bar{C}_i(m)$ is 24, that is 8 data blocks. In these 8 blocks, sensor $i$ samples 8 observation data in total. The resulted stacked observation matrix is

$$\bar{C}_i(m) = \begin{cases} C_i \cdot [e_3; e_6] & m = 3j-2 \\ C_i \cdot [e_1; e_4; e_7] & m = 3j-1 \\ C_i \cdot [e_2; e_5; e_8] & m = 3j \end{cases} \tag{48}$$

in the $m = 3j-2$ block, sensor $i$ sampled 2 observation data; in the $m = 3j-1$ and $m = 3j$ block sensor $i$ sampled 3 observation data.

More generally, for $M = 2^q$ and $d_i$ without $u \in N$ s.t. $d_i = 2^u$ , the structure of $\bar{C}_i(m)$ is

$$\bar{C}_i(m) = \begin{bmatrix} C_i \cdot e(d_i - r_i(m)) \\ C_i \cdot e(2d_i - r_i(m)) \\ \vdots \\ C_i \cdot e(S_i(m)d_i - r_i(m)) \end{bmatrix} \tag{49}$$

where $S_i(m)$ is the number of matrix rows which is the maximum integer s.t.

$$[S_i(m)d_i - r_i(m)] \le mM$$

# 5 SIMULATION

This section gives the simulation of the algorithm developed in this paper to demonstrate its validity. Multiscale sequential fusion result of 3 sensors whose sampling interval are respectively 1,2 and 3 are compared with that of one single sensor using Kalman filter method.

The parameters used in the simulation are $A(k) = 0.96$ , $Q(k) = 1$ , the initial state is $x_0 = 1$ , $P_0 = 1$ .Stacking the system with block length $M = 4$ , then use the Haar wavelet to implement wavelet transform. The observation parameters are $\bar{C}_1(m) = I_4$ and $R_1 = 0.5$ ; $\bar{C}_2(m) = 0.5 \cdot [0\ 1\ 0\ 0; 0\ 0\ 0\ 1]$ and $R_2 = 0.1$ ;

$$\overline{C}_3(m)=\begin{cases}2\cdot[0\ 0\ 1\ 0] & m=3j-2\\ 2\cdot[0\ 1\ 0\ 0] & m=3j-1\\ 2\cdot[1\ 0\ 0\ 0;0\ 0\ 0\ 1] & m=3j\end{cases},$$

$R_3=0.01$.

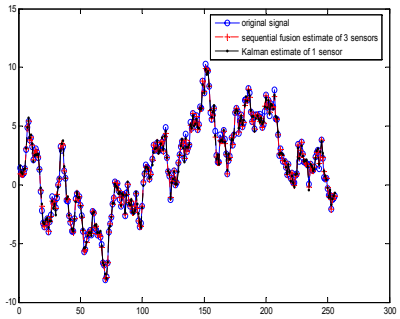



Figure 3: Sequential fusion result via single sensor estimate.

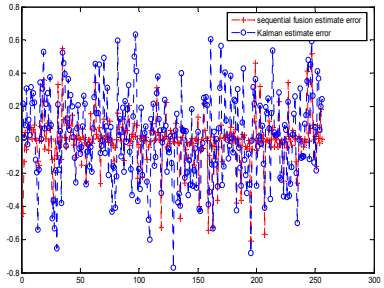



Figure 4: Sequential fusion error via single sensor estimate error.

It is easy to see that using the method mentioned in section 4 to design the stacked observation model can realize the multiscale sequential fusion of non-$2^n$ sampling data. Compare the fusion estimate using this multiscale sequential fusion method and one single sensor estimate using Kalman filter,we conclude that it is an efficient method to process fusion problem with non-$2^n$ sampling observation data, which is an obstacle of multi-scale data fusion.

The mean of absolute error (MAE) displayed in table 1 compare the estimate error accuracy based on one single sensor 1 using Kalman filter method and that based on sensor multi-sensor using the multiscale sequential fusion. We find that the estimation accuracy improved 2.53 times.

Table 1: MAE of sequential fusion and single KF.

| | MAE |
|---|---|
| single sensor Kalman filter | 0.2169 |
| 3 sensor sequential fusion | 0.0857 |

# 6 CONCLUSIONS

Hybrid wavelet-Kalman filter method can obtain the real time multi-scale estimate of dynamic system. The multiscale sequential fusion algorithm based on it can easily fuse information from multiple sensors sequentially without designing other complex fusion algorithm. In addition, the hybrid wavelet-Kalman filter multiscale sequential fusion method can used to fuse non-$2^n$ sampled data just by designing the periodically varied stacked observation matrix.

# ACKNOWLEDGEMENTS

This paper is supported by NSFC (60434020, 60572051); the Education Key Project (07ZZ102) and the Education Development Project (08YZ109) from Shanghai Municipal Education Commission.

# REFERENCES

Wen Chenglin, Zhou Donghua, 2002. *Multi-scale estimate theory and application*, Being jing: Qstinghua Publication House.

Wen Chenglin, Jin Feng, Zhou Donghua, 2002. *Multi-scale estimate algorithm for one single sensor and one model*, Journal of Electronics 30(6):819-822.

Lang Hong, 1994. *Multi-resolution distributed filtering.* IEEE Transactions on AC, 39(4): 853-856.

Tong Xinzheng, A.A.Girgis, E.B.Makram, 2000. *Hybrid wavelet-Kalman filter method for load for forecasting.* Electric power systems research. 54(2):11-17.

Wen Chenglin, Xie Jin, Zhou Funa, Wen Chuanbo, 2006. *A New Hybrid Wavelet-Kalman Filter Method for the Estimation of Dynamic System* .Journal of Electronics(China),23(1):139-143.

Chenglin Wen, Chuanbo Wen, 2006. *The Multiscale Sequential Filter with Multisensor Data Fusion.* Proceedings of the 25th Chinese control conference:483-488. Harbin, Heilongjian, China.

Funa Zhou, Tianhao Tang, Chenglin Wen, 2007. *A New Multiscale Estimation Scheme for Dynamic System.* Proceedings of the 26th Chinese control conference(5):396-400. Zhangjiajie, Hunan, China.

# MULTICHANNEL EMOTION ASSESSMENT FRAMEWORK
## *Positive and Negative Emotional Dichotomy*

Jorge Teixeira
*Faculdade de Engenharia da Universidade do Porto, Rua Dr. Roberto Frias s/n, Porto, Portugal*
*teixeira.jorge@fe.up.pt*

Vasco Vinhas, Eugenio Oliveira, Luis Paulo Reis
*Faculdade de Engenharia da Universidade do Porto, Rua Dr. Roberto Frias s/n, Porto, Portugal*
*LIACC - Artificial Intelligence and Computer Science Laboratory, Rua Campo Alegre 823, Porto, Portugal*
*vasco.vinhas@fe.up.pt, eco@fe.up.pt, lpreis@fe.up.pt*

Keywords: Medical Signal Acquisition, Data Analysis and Processing, Emotion Assessment, Electroencephalography, Galvanic Skin Response.

Abstract: While affective computing and the entertainment industry still maintain a substantial gap between themselves, biosignals are subject of digital acquisition through low budget technologic solutions at neglectable invasive levels preventing users from focusing their awareness in the equipment. The integration of electroencephalography, galvanic skin response and oximeter in a multichannel framework constitutes an effort in the path to identify emotional states via biosignals expression. In order to induce and detect specific emotions, gender specific sessions were defined based on the International Affective Picture System and performed in a controlled environment. Results granted by distinct analysis techniques showed that high frequency EEG waves are strongly related to emotions and are a solid ground to perform accurate emotion classification. They have also given strong indications that females are more sensitive to emotion induction. On the other hand, one might conclude that the attained success levels concerning relating emotions to biosignals are extremely encouraging not only to the continuation of this research topic but also to the application of these results in domains such as multimedia entertainment, advertising and medical treatments.

## 1 INTRODUCTION

Affective computing is consistently becoming a confirmed scientific domain with practical applications while the entertainment industry as a whole, and specially the cinematographic and videogame branches, which have been closing the semantic gap between them, constitute an economic giant. Having this macrocontextualization in mind, the authors have already engaged a research project with the main intention of using emotion assessment through biosignals to promote both subconscious interaction and individual specific appropriated content delivery.

The presented study finds itself integrated in this scope, as it perfectly falls in the emotion assessment research module. The proposed system constitutes a solid technologic framework that intends to enable biological information acquisition in a controlled environment having as initial hypothesis the existence of human physical expression of emotional states that can be objectively measured by relatively inexpensive equipment. The multichannel structure was defined by exploiting known techniques, namely electroencephalography, galvanic skin response and heart rate monitoring, and emotional states were induced using third-party catalogued pictures.

The first goal was to effectively define, build and test an experimental session framework where subjects followed a given strict protocol in order to visualize and/or interact with multimedia content. This framework was designed not only to collect data but also to constitute a validation environment. The second objective was to, using the given platform, identify specific, controlled and extractable biological signals that could be used as emotion index factors applied to all subjects or to a characteristic group of equals.

The confirmation of the initial hypothesis and project goals would enable its immediate application in developing a full or semi-automatic emotion classification engine that would be able to apply and use the identified signal patterns to, in real-time, identify

the subject's emotional states with high accuracy.

In order to better detail the presented study, this document is structured as follows: the domain state of the art is described in the next section, in section 3 the multichannel emotion assessment framework is presented with special emphasis in the most significant decisions. Still in that section, results are presented and related conclusions are extracted in section 4 as well as are identified future work areas and practical domains of application are suggested.

# 2 STATE OF THE ART

The emotional state of human beings belongs to a complex theme since its definition is not unique and its essence not consensual. An overview of the emotion assessment is presented in the next subsection, as well as a brief description of the most common approaches to emotional induction, and finally a reference to equipment solutions.

## 2.1 Emotion Assessment

The emotion itself can be seen as a consequence of an action or an environment cause so that the induction of a specific emotional state is tightly connected with an arousal procedure. In order to identify and assess an emotion, patterns are used and they constitute different approaches to the emotional induction, which will be discussed in the next subsection. Apart from the induction, the classification is essential, and can be accomplished based on a coincidence of values on a strategic number of dimensions (Logothetis, 1957). Based on this study, the emotion assessment can be analyzed through three distinct dimensions. The two primary levels are the valence and the arousal, and the secondary one is the dominance, which has a weaker relationship with the others (P.J. Lang, 2005)(A. Mehrabian, 1974).

In order to best analyze the assessment of the pictures, it is generally used an affective space. This is a standardized method to graphically display the emotional assessment results of the pictures. According to the valence and arousal mean values, it is plotted a bidimensional graph where the horizontal-axis represents the arousal and the vertical-axis the valence, both scaled from 1 to 9.

## 2.2 Generic Approach to Emotion Induction

There is not a process for emotional induction that is perfectly suitable for all cases, but a group of different approaches to achieve the same objective. A prevalent method to induce emotional processes consists of asking an actor to feel or express a particular mood. This strategy has been widely used for emotion assessment from facial expressions and to some extent from physiological signals (G. Chanel, 2005). However, even for expert actors for whom the capacity to achieve a specific emotional state is obvious, it is hard to guarantee that the physiological responses are consistent and reproducible by other non-actor people.

An alternative approach to the emotional induction is composed by multimedia stimuli. Music, images, videos and video-games belongs to a category of stimuli that has significant advantages compared with the induction through actors, since there is no need of actors and the quality of the induced emotions is greater as they are more realistic.

## 2.3 Equipment Solutions

Emotions' assessment needs reliable and accurate communications with the subject so that the results are conclusive and the emotions correctly classified. This communication can occur through several channels and is supported by specific equipment. The BCI - Brain Computer Interface - is directly connected to this area and uses two different approaches, invasive and non-invasive methods. The invasive methods are clearly more precise, however more dangerous and will not be considered for this study. On the other hand, non invasive methods such as EEG, fMRI, GSR, oximeter and others have shorten the distance between the utopia and the truth of understanding the human brain behaviour, gathering together the advantages of inexpensive equipment and non-medical environments.

Due to the medical community skepticism, EEG, in clinical use, it is considered a *gross correlate of brain activity* (Ebersole, 2002). In spite of this reality, recent medical research studies (Pascalis, 1998)(Aftanas, 1997) have been trying to revert this scenario by suggesting that increased cortical dynamics, up to a certain level, are probably necessary for emotion functioning and by relating EEG activity and heart rate during recall of emotional events. Similar efforts, but using invasive technology like Electrocorticography (ECoG), have enable complex BCI like playing a videogame or operating a robot (Leuthardt, 2004).

Some more recent studies have successfully used just EEG information for emotion assessment (K. Ishino, 2003). These approaches have the great advantage of being based on non-invasive solutions, enabling its usage in general population in a non-medical environment. Encouraged by these results,

the current research direction seems to be the addition of other inexpensive, non-invasive hardware to the equation. Practical examples of this are the introduction of GSR and oximeters by Takahashi (Takahashi, 2004) and Chanel et al(G. Chanel, 2005).

On this study three non-invasive equipments will be used in parallel so that the reliability of all the procedures is guaranteed. A Neurobit Lite EEG device with one active electrode and two references, a Thoughtstream biofeedback system Galvanic Skin Response with two dry electrodes and an oximeter with a finger sensor.

# 3 PROJECT DESCRIPTION AND RESULTS

In this section it will be given a brief overview of the whole project development.The procedures and methods involved on this study were grouped into two parts. The first one deals with the emotional induction approach used on this study and the last one reveals the experimental conditions used along the experimental sessions.

A good experimental control and an easy, yet efficient, method for results comparison are key factors that demand an effective set of visual stimuli. The IAPS library is so the most indicated emotional induction method, as it has been widely used through the research community and all the pictures classified according to valence, arousal and dominance (L. Aftanas, 2001)(G. Chanel, 2005)(M. Muller, 1999).

The experimental conditions are an essential issue to the validation and acceptance of the results obtained.

A total of twenty eight subjects, seventeen males and eleven females, all right-handed aged eighteen-thirty years old took part in this study. All subjects had access to an introductory text for the experimental session in order to access the essential information about the main procedures involved and a questionnaire filled before each session to avoid possible barriers as mental diseases.

The experimental results of this study are based on a statistical analysis.

This analysis indicates that mens behaviour is different from the womens one, since the mean amplitude of the high frequency brain waves is higher along the entire experimental session.

Considering the GSR data, it is presented the slope variation between the three stages - happiness, neutral and sadness - and analyzed its behaviour along the complete session.

The heart rate analysis indicates an unexpected and almost undetectable variation of its value.

# 4 CONCLUSIONS

In what regards to the first topic, one ought to assert that the initial study's main goals were completed undertaken. The described experiments achieved to develop and test a solid framework to conduct controlled emotionally-evocative experiments enabling flexible capability of recording and monitoring biosignals in real-time.

At a more detailed level, it was possible to define dynamic gender-designed emotional sessions, with fine tuning capabilities, in order to trigger specific emotional states. The data provided from the conducted experiments was input to the detailed system's architecture that proved to be able to supply sessions with biosignals real-time monitoring and storage for physical and temporal independent processing. The achieved results also demonstrated that the postsession data processing techniques were efficient and effective in what concerns to emotional states/biosignals correlation identification.

Having the study's results, presented in the previous section, as solid ground, it is possible to confirm that basic emotional states have biological manifestations capable of being captured and recorded by the selected equipments, specially with EEG and GSR techniques.

Concerning subject variables, it is plausible to state that females react more aggressively to the presented pictures, triggering with more expression and effectiveness the desired emotional states. These objective data was also corroborated by the interviews conducted at the end of each sessions, where female subjects consistently stated that they felt happy and in a good mood in the first stage of the session and sad at the end. In the same interviews, a high percentage of male subjects affirmed that they did not felt a deep emotional commitment along the picture presentation.

The fact that gender is a key factor in what concerns emotional state triggering through multimedia content constitutes the first major contribution of the present study.

Exploiting the evidence that EEG signals were strongly influenced by the subject's emotional states, a more detailed data analysis was performed, as illustrated in the previous section, with special focus to high frequency signals, namely beta and gamma waves. The data provided strongly suggests that, specially in female subjects, high frequency relative

EEG values are directly correlated to valence, independently of their initial standard level. The data provided suggests that beta and gamma waves strongly seam to vary directly with valence, enabling, indirectly and with conjugation with other inputs, emotional state detection.

Despite the described positive outcome, there were identified some features that, although did not match the initial assumptions, have already been subject of turnaround strategy definition. The first one is related to the inexistence of generic significative changes in heart rate values along experimental sessions and the second resides in the fact that the expected GSR readings curve – high conductivity with high arousal situations – was not recorded as often as predicted. The authors believe that the existence of these issues is grounded on the fact that the designed emotionally-evocative sessions based only on pictures do not trigger such strong emotions capable of significantly influence biosignals such as heart rate. It is believed that it would be necessary much strong multimedia content provided in a more immersive environment so that subjects could be more deeply involved.

### 4.1 Future Work

The main future work topics are not only related to this particular study, once it is a spin off/module of a major one, but also with the main global project. With this in mind, there were identified the following areas:

- More Sophisticated Equipment Reinforcement: It is intended to acquire more sophisticated equipment, specially and specifically in what concerns to a multi-channel EEG and a more sensible and reliable GSR;
- Equipment Diversity: It would be useful to integrate in the developed framework new equipments capable of reading and extract more biosignals, namely pupil dilatation, voice analysis and facial expression recognition;
- More Detailed Emotion Classification: using the depicted key factors with conjugation with others provided by the study continuation and data volume and diversity enhancement brought, by new equipment acquisition, it would be plausible to perform automatic subject emotion classification with deeper detail levels.
- Software Control: The accomplishment of the previous items would enable both conscious and subconscious control of several tools and/or multimedia contents;

Considering the studied problem as a whole, specifically the emotion classification topic, several practical domain applications are not only feasible but also attractive. Most of the immediate technology adaptations shall reside in the entertainment industry, both in audiovisual and videogame branches through multimedia content adaptability to user's emotional states. Other possible application areas are user interface enhancement, direct advertising and medical applications, namely in phobia treatments and psychological evaluations.

## REFERENCES

A. Mehrabian, J. R. (1974). An approach to environmental psychology. In *The MIT Press*.

Aftanas, L. (1997). Nonlinear forecasting measurements of the human eeg during evoked emotions. In *Brain Topography*, volume 10, pages 155–162.

Ebersole, J. (2002). *Current Practice of Clinical Electroencephalography*. Lippincott Williams & Wilkins.

G. Chanel, J. Kronegg, D. G. (2005). Emotion assessment: Arousal evalutation using eeg's and peripheral physiological signals. In *Technical Report*.

J. Allen, J. K. (2006). Frontal eeg asymmetry, emotion, and phychopathology: the first, and the next 25 years. In *Biological Psychology*, volume 67, pages 1–5.

K. Ishino, M. H. (2003). A feeling estimation system using a simple electroencephalograph. In *Proceedings of 2003 IEEE Int. Conference on Systems, Man, and Cybernetics*, pages 4204–4209.

L. Aftanas, A. A. (2001). Time-dependent cortical asymmetries induced by emotional arousal: Eeg analysis of event-related synchronization and desynchronization in individually defined frequency bands. In *Int. Journal of Psychophysiology*, volume 44, pages 67–82.

Leuthardt, E. (2004). A braincomputer interface using electrocorticographic signals in humans. In *Journal of Neural Engineering*, pages 63–71.

Logothetis, N. (1957). The measurement of meaning. In *University of Illinois Press*.

M. Muller, A. K. (1999). Processing of affective pictures modulates right-hemispheric gamma band eeg activity. In *Clinical Neurophysiology*, volume 110, pages 1913–1920.

Pascalis, V. D. (1998). Eeg activity and heart rate during recall of emotional events in hypnosis: relationships with hypnotizability and suggestibility. In *Int. Journal of Psychophysiology*, volume 29, pages 255–275.

P.J. Lang, M. Bradley, B. C. (2005). International affective picture system (iaps): Affective ratings of pictures and instruction manual. In *Technical Report*.

Takahashi, K. (2004). Remarks on emotion recognition from bio-potencial signals. In *The second Int. Conference on Autonomous Robots and Agents*.

# MODEL BASED DESIGN OF NETWORKED EMBEDDED SYSTEMS
## *A Modeling Approach using FlexRay as an Example*

Johannes Klöckner, Sven Köhler and Wolfgang Fengler
*Institute of Computer Engineering, TU Ilmenau, Ilmenau, Germany*
*johannes.kloeckner@tu-ilmenau.de, sven.koehler@tu-ilmenau.de, wolfgang.fengler@tu-ilmenau.de*

Keywords: Model Based Design, FlexRay, CAN, MLDesigner, Network Simulation, Building Blocks, Fieldbuses.

Abstract: This paper presents a work in progress on a method to create system level models of networked systems in automotive applications. It introduces an example, that shows a strategy to create models, providing high flexibility in terms of interoperability, field of application, reusability and replaceability. The chosen modeling tool contains a multi-domain simulator and allows a mission and system level design. Beside the exposition of the basic architecture of the model there is a description of various model parts showing the variety of different levels of abstraction. The grade of reuseability of the developed building blocks is very high. Finally a perspective for future extensions towards a general modeling strategy for various networked applications in embedded systems is provided.

## 1 INTRODUCTION

In recent years the complexity of embedded systems has been growing to large, hard to manage dimensions. Additional to the system design itself the networking of embedded systems becomes more complicated and requires a lot of planning and design decisions. In contrast to that growing complexity the time to market dramatically decreases and development costs need to be reduced to be competitive.

At present the tool and method support for efficient top-down development processes is insufficient. One way to deal with this problem is the model driven development of systems. The creation of models eases the hard- and software development and creates documented interfaces. Some concepts permit to automatically produce source code to enhance a rapid development. Most modeling technologies allow a simulation to validate behavior of the modeled system or network without the demand to build an expensive prototype. Furthermore, these techniques provide technologies for performance tests, thus supporting e.g. the optimization of applications.

One field of technology, where models can be used to accelerate the development, is the automotive industry. Today's vehicles contain a large amount of electronic systems, the variety ranges from driver assistance to passenger entertainment. These established features and new applications like drive-by-wire or networked cars[1] increase the need for new networking technologies, that offer fast and reliable time critical communication, as well as the design of a complex complete system.

FlexRay (FlexRay Consortium, 2007) is a new communication system, that offers real time features as well as high bandwidth by the use of a flexible time triggered system. The industry promotes this communication protocol as an important future technology and the migration to FlexRay has already begun. In this paper we will present a modeling strategy for networked embedded systems in automotive environment, that is best suited for the upcoming time triggered communication systems providing real time capability and high performance. The FlexRay protocol will serve as the central example for the developed approach.

This paper is organized as follows. Section 2 describes the state of the art in system development. Section 3 introduces the basic modeling strategy. Section 4 presents the selected modeling tool. Section 5 describes the different model elements. Section 6 presents the drawn conclusions. Finally, Section 7 gives a brief overview of the next development steps.

[1] Car-2-Car communication

## 2 STATE OF THE ART

A lot of tools support the configuration and development of FlexRay systems. Mostly they are enhancements of already known approaches deployed for the design of *controler area network* (CAN) (Robert Bosch GmbH, 1991) based systems, e.g. the tool set of the company Vector Informatik GmbH[2] (Carsten Böke, 2006). The tools are principially used for application development. Simulation and monitoring is available in conjunction with a hardware node. With regard to model based design and simulation, which allows system design and analysis, these tools are not suitable. MATLAB (The MathWorks, 2007) is a system design tool focused on continuous time models, but it is not well suited to build a model for discrete event simulations, like communication protocols. The company DECOMSYS[3] provides an extension block to use FlexRay inside MATLAB. This precast block allows a hardware based model simulation. Due to this and the limited block access the analysis varieties are restricted. Current research on schedulability analysis (Richter, 2007) is based on stand alone solutions. They are not integrated into the development process of communication systems.

There is no adequate combination of system analysis and system design. Within a model based design an efficient development of a complete system is possible (Salzwedel, 2004). A system is a composition of different building blocks. The building blocks shall be grouped into categories, e.g. communication protocols. Each category should offer common interfaces to provide a high grade of reusability and exchangeability. Also a category has to be divided in different parts containing different realizations of a system element, which possess variable levels of abstraction. To support the system design and analysis mechanisms are needed to enable the monitoring of the system and communication behavior, the fault injection and to allow an easy configuration of the system.

The tool MLDesigner (MLDesign Technologies Inc., 2007) fulfills the requirements of a model based design allowing system analysis and development. It will be described in section 4.

## 3 MODELING STRATEGY

The modeling approach introduces a strategy to support a generalized model based top-down development process focused on networked embedded systems in automotive applications. A networked system comprises several components. The predefinition of common interfaces supports the modeling approach and allows an easy exchange of components. The components themselves can be seen as small systems composed of basic elements, e.g. the application, the operation system and the communication protocol. In a top-down development process compatibility and reusability can be achieved by using building blocks. Basically the system is divided into two parts, application and communication. This partitioning is equivalent to a division between function and architecture. An important element in networked systems is the communication architecture. It is necessary to compare different network topologies and determine the influence of the communication type on the system design. To allow an easy comparison between variable protocols, building blocks provide reusability and exchangeability. Due to this the initial focus can be set to the communication structure and building blocks on the level of communication protocols. FlexRay is selected as first exemplary realization.

To address as many use cases as possible, different levels of abstraction are required. For a given system there are various relevant examination aspects, the efficient simulation of which requires the use of particular models and building blocks. With building blocks a complete FlexRay system can be modeled on an abstract level to simulate the high-level system behavior. In terms of an analysis of the detailed timing and synchronization behavior of a FlexRay node another more detailed model is provided. Both models include options for fault injection and monitoring. To ensure a maximum of reuseability of the modeled components a common interface between both abstract and detailed model components is specified. The models are organized in libraries and provide different implementations of system elements. These system elements are, e.g. different applications related to different abstraction levels and also different communication protocols.

For future work the interface towards the application or other so called upper layers is very important to allow the integration of further protocols. These protocols have to implement the interface in a similar way to achieve a maximum of reuseability and exchangeability of the system elements. The support of a multiplicity of protocols allows a comparison of different communication systems in a particular scenario. Also the analysis of heterogeneous networks and the analysis of a structural migration is possible. These tasks are relevant regarding FlexRay and CAN.

[2] http://www.vector-worldwide.com/

[3] http://www.decomsys.com/

# 4 MODELING TOOL

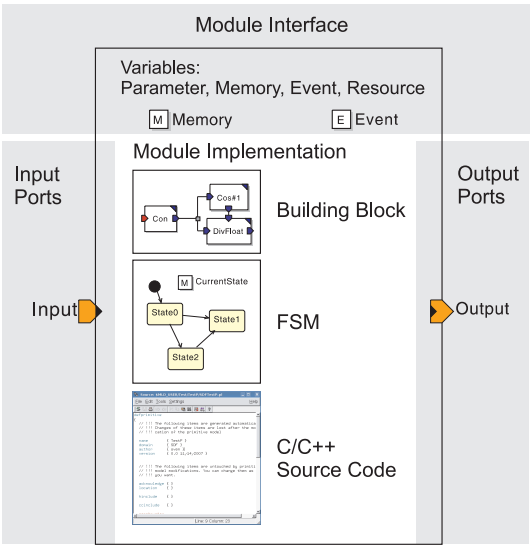



Figure 1: Model structure of a basic communication system.

The tool MLDesigner by MLDesign Technologies, Inc. is dedicated to improve the design process from early concepts to implementation. It is a tool offering mission and system level design, including operational, architectural and functional level, and evaluation facilities. Build upon the well known Ptolemy project of UC Berkeley (The Ptolemy Project, 2007) it offers the same modeling techniques, but extends them with new models and a better graphical representation. Like Ptolemy, MLDesigner provides different models of computation as so called domains. It offers a *discrete event domain* (DE) and *finite state machines* (FSM), *synchronous data flow domain* (SDF) as well as a *continous time/discrete event domain* (CTDE) for numerically solving models given as differential equations. The models created in MLDesigner are structured similar to those of Ptolemy. The top level of the modeling hierarchy is called *system*. A *system* includes all other used blocks and has no interface to other blocks outside. Examples for building blocks are shown in Figure 1. Those can be *primitives* like FSM or C/C++ source code as elementary blocks in the hierarchy. Other levels in the hierarchy are formed of *modules*, intern containing *modules* or *primitives*. All building blocks possess different types of variables. *Parameters* cannot be altered during simulation meanwhile *memories* depicted by an ”M” in a box shown in Figure 1 are changeable. *Variables* can either be local or linked to a similar element in the next higher hierarchy level. Building blocks can communicate with their environment by using these linked *variables* or through *ports* shown as arrows on the bounding box of a building block. By the use of *wormholes* the communication between building blocks of different domains is supported. This very flexible and powerful modeling paradigm offering a top down design makes MLDesigner the best suited tool for our modeling strategy.

# 5 FLEXRAY LIBRARY

## 5.1 Basic Model Structure

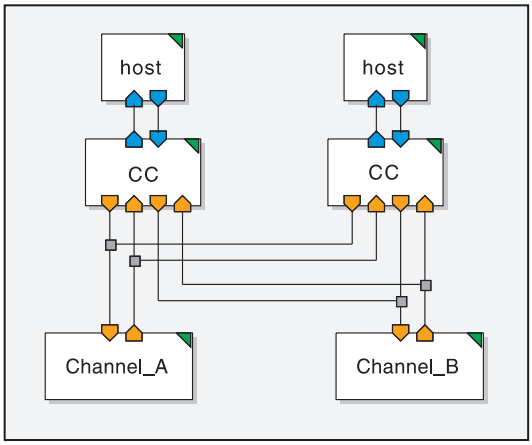



Figure 2: Model structure of a basic communication system.

Each communication system shown in Figure 2 is a composition of three different elements: *host*, *communication controller* (CC) and *channel*. A combination of a *host* and a CC is called *node*. The *host* contains the application and is responsible for the configuration of the CC, initiation of the sending operation and processing of the received data. At this point the division into parts 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. Each cluster consists of two channels, *channel A* and *channel B*. The CC itself implements the communication protocol, e.g. the frame transmission and the frame reception, and contains memories, which represent the controller state and configuration data.

Interaction between model elements is realized using different data structures as signals. For the signaling between *host* and CC the communication is service based. Referring to this the data structure contains a service identifier, a sub service identifier and additional data required by the selected service. The definition of these services is based on the FlexRay specification. Dependent on the model type, abstract

or detailed, the data exchange between different nodes via the channels is frame based or on bit level. To allow an easy exchange of the two models, abstract and detailed model provide the same interface to the *host*. This so called *controller host interface* (CHI) is part of the CC. In the FlexRay specification the interface functions are only roughly described, therefore this basic description has to be filled to realize a precise model.

In the following the models are described in more detail. First there will be a short description of the CHI, which is an important part of the abstract and the detailed model, because it provides the same interface towards an upper layer. Afterwards the abstract and the detailed CC models will be explained.

## 5.2 Model Elements

### 5.2.1 Controller Host Interface

The CHI is responsible for the data and control flow between CC and *host*. Beside the function as interface, the CHI administrates the transmission and reception buffer, manages the reception filter and provides access to configuration and status data. Both reception and transmission buffer are CHI local memories, the same applies for the reception filter. The receive and transmission buffers are implemented with unrestricted capacity. With regards to the primary aims of the library it is not necessary to take limited buffers into account. Each buffer is realized as a vector of data elements containing the relevant information, example given frame identifier, channel, payload length and payload data.

The CHI has two different interfaces, one to the host by using the mentioned service based data structure and a second interface towards the CC protocol functions. This second data structure can be interpreted as purely signal based. It is derived from the protocol's internal communication and contains a signal identifier and a field for additional data. A *host* has write access to the transmission buffer and read access to the reception buffer by using CHI services. On the other side the protocol can use signals to write received data into the reception buffer and to retrieve data to send from the transmission buffer.

Each CC contains its own configuration, these communication parameters are realized as memories, too. Write access to the memories is only possible by using the CHI and the provided services. Each host is responsible for the correct configuration of its own CC. So for configuration aspects it is not necessary to parameterize the CHI module. To support the creation of a *host* an additional library is provided allowing initialization, configuration, message sending and reception.

### 5.2.2 Abstract CC Model

The design of a model starts with the question: What is the operational aim of the model? Creating a detailed model the answer is easy, the model has to be built as accurately as possible. To achieve this the specification is used as blueprint. As mentioned before the abstract model should allow a system analysis or development of systems on a higher level, e.g. to support decisions in an early stage of development. The abstract model of the FlexRay CC uses some simplifications concerning the clock synchronization mechanism, the temporal behavior and the frame based data transmission. An external central time master called *Global Clock* (GC) is responsible for synchronization and the timing of the FlexRay communication.

The GC generates the cluster wide valid time, which is represented by the so called *macroticks* (MT). Important protocol values like the *slot counter* for both FlexRay channels and the *cycle counter* are derived from the MT. Each controller needs information about important time events. In a FlexRay cluster these are the slot starts. The GC announce a change of the *slot counter* to all controllers. A controller retrieves the actual counter values and checks, if valid transmission data is available. In addition to the medium access control the protocol model is also responsible for data transmission and reception.

The description of the model can be divided into two parts, one is the interface and the other is the functionality and structure. Basic information about the interfaces are already given, on one side of the controller it is the well known CHI and on the other side the controller communicates with the channel via a data structure. This channel data structure contains all relevant data: frame identifier, payload data length, cycle count, payload data and also some data for administrative tasks, e.g. an error indicator. The abstract model consists of the following modules: CHI, *UpdateStatus*, *SendtoChannel* and *ProcessRecData*. As shown in Figure 3 the controller has two modules *SendtoChannel* and two modules *ProcessRecData*, one connected to *channel A* and one connected to *channel B*. Each module itself has a complex internal structure consisting of further blocks based on modules, MLDesigner primitives and newly developed custom primitives. The internal structure of the modules will not be discussed in detail.

The module *UpdateStatus* receives events generated by the GC, updates the local time and triggers the module *SendtoChannel* when a new slot starts. This

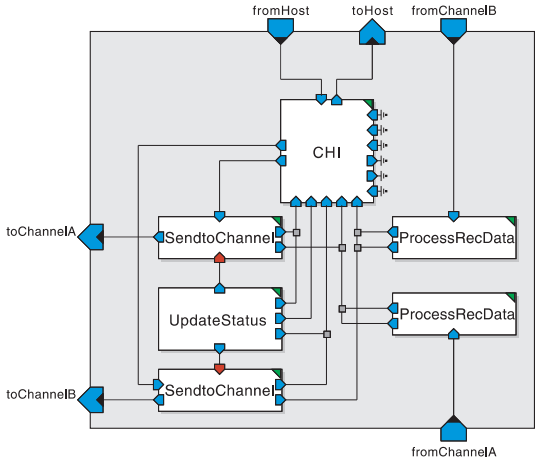



Figure 3: Internal structure of the abstract CC model.

module is responsible for the medium access control. A data request signal is send to the CHI. Dependent on the reply signal a data frame is created and send to the channel. This data frame is forwarded by the channel to all connected controllers and received by the module *ProcessRecData*. After frame reception and conversion the received data is forwarded to the CHI containing the reception buffer.

To associate the simulation with the reality there is a coherence between real time and simulation time. An integer time step in the simulation is equivalent to a second. This assumption is compatible to the detailed model.

Preliminary performance tests have shown a real-time to simulation time ratio of 1:200 in case of the abstract model. Some small changes improved the ratio to 1:20. It can be assumed that further improvements are possible, example given by using optimized data representation.

### 5.2.3 Detailed CC Model

The detailed model should allow an exact analysis of the protocol and a system. The main benefit is, that the internal behavior and the internal protocol mechanisms can be visualized. As a result of the detailed implementation of the clock synchronization and the time generation mechanisms there are additional functionalities feasible, like the simulation based optimization of configuration parameters.

The model is called detailed, but there is also a level of abstraction by comparison to a real system. Beside the memory management the abstraction concerns the communication. In real systems the data is transmitted using analogous signals, in context of the detailed model the communication is bit based. Each controller transmits and receives a bit string and interprets the data. An advantage of the modeling tool MLDesigner is, that the model can be easily enhanced using the CTDE domain, if a more detailed model processing an analogous signal waveform is needed.

As mentioned above the FlexRay specification is used as a kind of blueprint for the detailed model. The specification itself is divided in different parts: the protocol (FlexRay Consortium, 2005b) and the electrical physical layer (FlexRay Consortium, 2005a) specification. Due to the bit based communication the focus lies on the protocol specification. The protocol is specified in a semi-formal way using text and SDL[4](ITU-T, 2002) to describe the functionalities. The SDL semantic has been implemented in a suitable way by using MLDesigner elements. The FSM domain in connection with DE modules allows the realization of the protocol analog to the SDL specification. In the FlexRay specification is some space left for interpretation. This concerns the realization and use of the CHI, the controller configuration and the memory management and the signaling and communication of the SDL processes. First to mention is, that all SDL processes are realized as FSMs. For the intercommunication of the FSMs the MLDesigner signaling concept is used in the following way. Analog to the aforementioned signal mechanisms a data structure is used. Different kinds of signals are used in the specification: pure signals, signals with data and function calls concerning the CHI. All of these are implemented with one data structure containing a signal name as identifier and a field to place additional data. The internal communication is realized by using this data structure. For the communication to other elements the detailed model provides almost identical interfaces in comparison to the abstract model. To the host the already known CHI interface is used. An important part is the configuration of the controller. According to the abstract model the configuration is initiated by the host and the design is supported with an additional library. Interaction with the channel element is performed with 0 and 1 as bit values. The CC model as shown in Figure 4 consists of a CHI module and different modules which are equivalent to the SDL processes defined in the FlexRay specification. The modules CSP (*Clock Synchonization Processing*), MTG (*Macrotick Generation*) and POC (*Protocol Operation Control*) exist only once, whereas the modules CODEC (*Coding/Decoding Processes*), CSS (*Clock Synchronization Startup*), FSP (*Frame and Symbol Processing*) and MAC (*Medium Access Control*) exist twice, one per channel. Each process is implemented as a finite state machine.

[4]Specification and Description Language

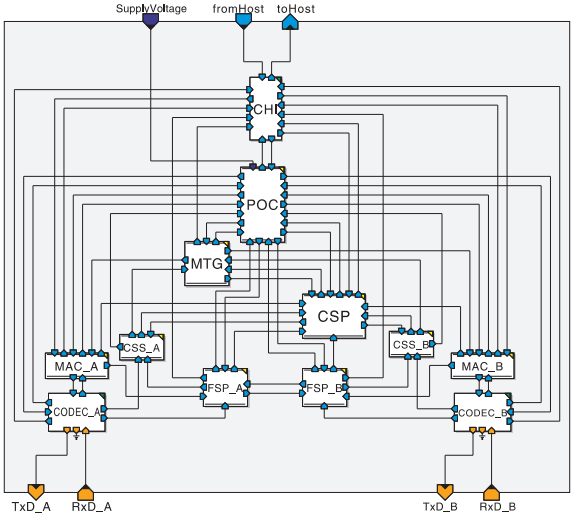



Figure 4: Internal structure of the detailed CC model.

#### 5.2.4 Channel

There are two different types of channels, one for an abstract CC and one for a detailed CC. Both have functions to inject faults, forward and delay data.

A delay for frames is necessary to model a correct timing behavior, since the frame data structure is completely transmitted at the start of a slot. Without a delay the frame would be also received at the start of a slot. Insertion of a delay, which is based on the transmission rate and the data length, establishes a correct timing.

Including errors and faults in a model is an important facet when analyzing a system. A great advantage of a model is the possibility to stress a system with an arbitrary fault. There is no need for complex fault injection or fault generation by using real scenarios with expensive hardware. The design of a fault model is a relevant part of the whole model library. Up to now there exist some basic mechanisms to inject faults, which will be improved in future. One fault injection mechanism is a node failure - a node loses bus synchronization and has to reconnect to the bus. Another important failure injection mechanism is the sending of erroneous data frames. The invalidation of an abstract data frame is split into two parts. A frame can be signed as invalid by using the error indicator. The selection, if a frame is faulty, is also based on a probability distribution and can be seen as the first part of the fault model. A second part is the selection of the error cause, this is probability controlled, too. The fault model, which is part of the detailed channel, allows a more extensive and sophisticated analysis. Not a whole frame can be marked as invalid, now one single bit can flip with the effect, that the error check mechanisms can be proofed as well. With an additional marking of flipped bits it can be tested, if a transmission error was detected or remaind undiscovered. Additional to these data frame errors it is possible to insert synchronization faults. In many cases fault models perform an important role, not only in the analysis of protocol performance, but also in the analysis of higher-level behavior. This could be the comparison of different fault handling strategies.

### 5.3 Validation

An important task in building models has not been mentioned yet, the validation of the model. Is the behavior of the developed model equivalent to the specified system? A short discussion of the validation of both models is given. The validation of the models is delicate because of the different monitoring feasibility. A real system allows only limited access to the internal behavior of the controller. As a result an exact comparison was not possible. The model allows a detailed monitoring, fault injection and analysis of every internal signal, none of which is provided by a real controller. In case of the not visible aspects the validation is based on the assumed behavior described in the FlexRay specification. The abstract model was tested by comparing it with the send behavior of a real system. As reference a system of two communicating FlexRay nodes is used. The communication and the timing behavior in normal and fault scenarios is compared with th results of an equivalent model. The results show, that the modeled system behaves like the real system. To validate the detailed model and the CHI another way is chosen. Here the model is compared to the specification and the behavior is shown by simulation of test scenarios. The complexity of the examples varies from small module tests to whole system tests. The detailed model was tested against the real two-node system, too. Comparison of the observable behavior of model and real system showed the same results.

## 6 CONCLUSIONS

In this paper a modeling strategy was described, which enables the design of models dealing with automotive communication systems like CAN and FlexRay. The resulting model provides concepts to monitor both, the behavior of networked systems and the internal behavior of the communication. This will allow easy fault injection as well as a schedulability analysis dependent on the selected bus system. The presented library developed for MLDesigner uses the *discrete event domain* in combination with FSMs and

demonstrates the approach with FlexRay as an example. Modularized components are part of the library. These so called building blocks allow an easy construction of a wide range of systems. The library includes modules implementing the FlexRay communication controller and a basic host system, which provides elementary functionalities needed in the design of hosts and applications. The components can be used to construct models for networked systems and subsystems or to develop models for gateways. The use of different communication protocols and gateways enable the simulation and the analysis of complex systems. Furthermore the concept is designed to assist in migrating parts of networked systems from CAN bus usage to FlexRay. The developed models with different degrees of abstraction support the development in early stages and enable the evaluation and verification of system properties prior to hardware design.

# 7 FUTURE WORK

The next step in applying this approach will be the design of monitoring modules to support the analysis of the FlexRay protocol itself during a simulation run and to provide all necessary information for debugging systems or generating results, e.g. for schedulability.

The completion of a CAN bus model is necessary. This will allow the seamless integration of CAN models into systems designed for FlexRay and vice versa, also simulating the migration of network parts from CAN to FlexRay will be possible.

Future work will also deal with the automated generation of models. Therefor the FIBEX (ASAM, 2007) standard will be utilized. It will allow to describe all important parameters of FlexRay and CAN systems to generate the network structure. With an extension of the FIBEX XML files might be possible to generate a complete system model based upon a FIBEX description using additional information. This process will be done by an XSLT transformation of the FIBEX file into an MLDesigner model also specified using XML files.

After the completion of the communication part of a system the functional part has to be added. Different modules implementing e.g. a gateway functionality will be designed. A future publication will deal with different mapping strategies usable within these gateways based on the concept and the models described in this paper. Modules for different host systems and operating systems will be added to support OSEK (OSEK/VDX, 2005) and OSEKtime (OSEK/VDX, 2001) compatible systems.

Finally we will create a system level model for an existing real world scenario in automotive applications to demonstrate our new approach in modeling systems.

# ACKNOWLEDGEMENTS

This work was funded by the ”Thüringer Aufbaubank”[5] under joint projekt number ”2006 VF 0014”.

# REFERENCES

ASAM (2007). *FIBEX - Field Bus Exchange Format Version 2.0.1*.

Carsten Böke (2006). Regler real testen. *Design/Elektronik 07/2006*.

FlexRay Consortium (2005a). *FlexRay Communications Systems - Electrical Physical Layer Specification Version 2.1*.

FlexRay Consortium (2005b). *FlexRay Communications Systems - Protocol Specification Version 2.1*.

FlexRay Consortium (2007). `http://www.flexray.com`.

ITU-T (2002). *ITU-T Recomondation Z.100 (08/02): Specification and Description Language (SDL)*.

MLDesign Technologies Inc. (2007). *MLDesigner Documentation, Version 2.7*. `http://www.mldesigner.com/`.

OSEK/VDX (2001). *OSEK/VDX time triggered operating system Version 1.0*.

OSEK/VDX (2005). *OSEK/VDX Operating System Version 2.2.3*.

Richter, K. (2007). Scheduling analysis for flexray. In *KFZ-Entwicklerforum und FlexRay Solution Day 2007*. WEKA.

Robert Bosch GmbH (1991). *CAN Specification Version 2.0*.

Salzwedel, H. (2004). Design technology development towards mission level design. In *49. Internationales Wissenschaftliches Kolloquium IWK'2004*.

The MathWorks (2007). *Matlab 7 - Desktop Tools and Development Environment*. `http://www.mathworks.com/`.

The Ptolemy Project (2007). `http://ptolemy.eecs.berkeley.edu/`.

[5] `http://www.aufbaubank.de`

# MODELING AND ESTIMATION OF POLLUTANT EMISSIONS

El Hassane Brahmi, Lilianne Denis-Vidal, Zohra Cherfi, Nassim Boudaoud
and Ghislaine Joly-Blanchard
*University of Technology of Compiègne, BP 20 529, 60205 Compiègne, France*
{*brahmiel, denislil*}*@dma.utc.fr,* {*Zohra.Cherfi, Nassim.Boudaoud, Ghislaine.Joly-Blanchard*}*@utc.fr*

Keywords: Modeling, Combustion, Diesel Engine, Kriging method, Pollutant.

Abstract: The European laws lead to the increase of emission constraints. In order to take into account these constraints, automotive constructors are obliged to create more and more complex systems. This paper presents two stage approaches for the prediction of NOx (nitrogen oxide) emissions, which are based on an ordinary Kriging method. In the first stage, a reduction of data will take place by selecting signals with correlations studies and by using a fast Fourier transformation. In the second stage, the Kriging method is used to solve the problem of the estimation of NOx emissions under given conditions. Numerical results are presented and compared to highlight the effectiveness of the proposed methods.

## 1 INTRODUCTION

The diesel engine is an internal combustion engine. At each cycle during the intake stroke, the combustion chamber receives a mixture of air and vaporized fuel via the injector (their flows are measured and controlled). Afterwards fuel vapor and air are compressed and ignited.
The mixture air-fuel is not stoechiometric during the combustion process. The unfortunate consequence is the creation of pollutants. In order to limit this problem, the European laws increase the constraints on pollutant gas emissions.
The main aim is the minimization of the NOx emissions under some constraints based on the Kriging model, by making a compromise with engine performance. In this case multi-objectives optimization will be considered. Then, it is necessary to simulate the pollutant behavior which is the subject of this paper.

A physical phenomenon model has been developed by S.Castric et al (S. Castric, 2007) in order to simulate the engine behavior. It takes into account the input parameters (fuel mass flow, air mass flow, exhaust gas recirculation ratio,...) and gives the corresponding state variables, particularly pressure, temperatures, fresh gas mass, mixed gas mass, and burned gas mass. It leads to a good representation of the experiment results. Strategies based on Lolimot (Local Linear Model Tree) and Zeldovich mechanisms (Heywood, 1988) have been developed in order to predict emissions of NOx (Castric, 2007). In the first case, the corresponding model can lead to singular points, which reduces the precision of the results. In the second case, the results are not satisfactory enough. On the other hand, the trend surfaces can be used, but it is difficult to go deeper with this approach because it consists of a classical regression based on the assumption of independence of observations, which is rarely checked with spatial data (S. Baillargeon, 2004).

Our choice is the Kriging method which takes into account the dependence of spatial data and has a variance that is minimal among estimators without bias. Moreover it leads to efficient results.

This paper is organized as follows: In the first section, the ordinary Kriging techniques are recalled. In the second section, two different approaches for modeling our problem are proposed. An efficient reduction model strategy is considered in order to apply the kriging method. Finally the Kriging method is applied to the reduced model. In the last section, numerical results are given followed by a short discussion.

## 2 ORDINARY KRIGING TECHNIQUES

Kriging methods is used frequently for spatial interpolation of soil properties. Kriging is a linear least squares estimation algorithm. It is a tool for interpo-

lation, which is to estimate the value of an unknown real function $F$ at a point $x_0^*$, given the values of a function $Z$ at some other points $x_1$,....$x_n$.

## 2.1 Ordinary Kriging

The ordinary Kriging estimator $\hat{Z}(x_0^*)$ is defined by:

$$\hat{Z}(x_0^*) = \sum_{i=1}^{n} \lambda_i Z(x_i). \quad (1)$$

where m is the number of surrounding observations $Z(x_i)$ and $\lambda_i$ is the weight of $Z(x_i)$. The weights should sum to unity in order to make the estimator unbiased. The weights are also determined such that the Kriging variance is minimal.
This leads to a classical optimization problem with equality constraint. The Lagrange multiplier theory is used in order to work out this problem. It gives a linear system to be solved (Davis.J.C, 1986)

## 2.2 Semivariogram

The semi-variogram is a function representing the spatial dependency, and has been obtained from the stationarity definition. It is based on the assumption of intrinsic stationarity for spatial data, the variation of a data set that is only dependent on distance $r$ between two locations where the variables values are $Z(x_i+h)$ and $Z(x_i)$ with $r=|h|$, can be given by the following semi variogram:

$$\hat{\gamma}(r) = \frac{1}{N(r)} \sum_{N(r)} [Z(x_i) - Z(x_j)]^2 \quad (2)$$

where

$$N(r) = \{(i,j) \quad tel\ que \quad |x_i - x_j| = r\} \quad (3)$$

where $N(r)$ is the pair number of $Z(x_i+h)$ and $Z(x_i)$ and $\hat{\gamma}(r)$ is the experimental semivariogram.

A variogram model should be fitted to such semivariogram. Different form of variogram model are available. In this study a power model was used:

$$\gamma(r) = C_0 + m.r^d \quad as \ \ h \geq 0, \ \ 0 < d < 2 \quad (4)$$

where $C_o$ is called the nugget effect, The least square method was used to estimate the parameters of experimental variogram and variogram models.

## 2.3 Cross-Validation

To evaluate the reliability of kriging estimation, cross-validation was used,and the mean square error (MSE) of the kriging-estimated values had been calculated.

The mean error ME is a measure of the estimation bias, and it should be close to zero for unbiased methods.

# 3 DESCRIPTION OF TWO MODELS

S. Castric et all (Castric, 2007), have developed a physical model for modeling the engine behavior, in order to minimizing the NOx emessions. To do this, he has divided into two sub-model: The first is a physical model, making the link between the input parameters and state variables. The second study the impact of the latter on NOx, the second part was not completely done. In this work, we are inspired of this original idea to give the two modelings below.

## 3.1 First Modeling

The first one consist of studying the impact of input parameters on the NOx without taking into account the state variables. In this case, a model will be built by taking into consideration 8 input parameters like: pressure in the rail injection,the exhaust gas recirculation ratio..., with the corresponding value of the NOx flow.
The choice of these parameters was recommended by experts, and multiple regression to study the impact of these parameters on the NOx, was confirmed it.

## 3.2 Second Modeling

The second one consist of studying the impact of state variables on the NOx, which is tantamount to build a model that uses ten state variables like: cylinder low pressure, temperature in the cylinder...
which are each one represented by a vector of 1334 components and the corresponding value of NOx flow.

## 3.3 Model Reduction

The data of the first model can be directly used for applying the Kriging method. It is not the case for the second one. In the latter case, the data have to be reduced.

The reduction process begins by studying the different correlations between the state variables and their corresponding p-value. The criterion which has been chosen consists in testing the p-value: if it is inferior to 0.05 the correlation is considered significant.

This analysis allows us to retain two state variables only: the cylinder low pressure $P$ and the mixed gas temperature in the cylinder $T_e$.

In the second step, the number of components of the two remaining signals is reduced. It has been accomplished by using the discrete Fourier transform. The function fft of Matlab returns the discrete Fourier

transform (DFT) of a vector, computed with a fast Fourier transform (FFT) algorithm. After calculating the coefficients, a minimum number of these are retained. This number allows to reproduce the initial signal with a relative error of order $10^{-2}$, which is reached with only 40 Fourier coefficients. The reduction of the number of points of each signal is tantamount to minimizing the number of Fourier coefficients representing that signal.

The two retained signals, representing respectively the cylinder low pressure and the temperature of the mixed gas in the cylinder, have been reduced to a number of 40 Fourier coefficients. Each signal has been reconstructed from these 40 coefficients with an acceptable error.

The following table presents the relative error committed, for the reconstruction of the two signals from the 40 selected coefficients.

| Type of signal | relative error |
|---|---|
| Cylinder low pressure | 0.01 |
| Temperature of the mixed gas | 0.02 |

# 4 NOX ESTIMATION

## 4.1 Numerical Results using the First Model

This subsection will be devoted to the presentation of the numerical results obtained in the case of the first modelisation, more precisely we give the mathematical model used to adjust the experimental variogram and the corresponding graph.

The model used in this part is given by equation 5. where:
$C_0 = 9.909432.10^{-1}$, $m = 5.281263.10^{-8}$, and $d = 1.798734$

Figure 1 shows the experimental semi-variogram and the mathematical model which adjusts it. This model has the power form, without bearing and with a nugget effect $C_0$ . Several models were adjusted and then compared, it was difficult to select the better model by eye.The cross validation has facilitated the work. She allows us to select the one, that minimizes the mean square error, which is presented in this Figure.

| Type of Indice | Value |
|---|---|
| Mean Error | 0.1082633 |
| Mean square error | 11.23740 |

Finally the kriging model is obtained and Figure 3 illustrate the comparison between measured and simulated emissions of NOx by using this first model.

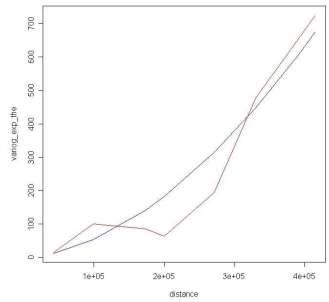



Figure 1: Experimental semivariogram and semivariogram model obtained using the input parameters.

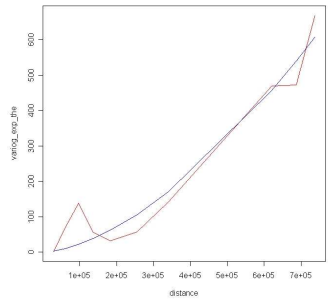



Figure 2: Experimental semivariogram and semivariogram model obtained using low pressure and temperature.

## 4.2 Numerical Results using the Second Model

This subsection is devoted to the presentation of the numerical results obtained in the case of the secon modelisation more precisely we give the mathematical model used to adjust the experimental variogram and the corresponding graph

The model used in this part is given by equation 5. where:
$C_0 = 9.917759.10^{-1}$, $m = 1.277732.10^{-}7$, and $d = 1.648926$.

Figure 2 shows the experimental semi-variogram and the mathematical model which adjusts it. This model has the power form, without bearing and with a nugget effect $C_0$ . Several models were adjusted and then compared, it was difficult to select the better model by eye. The cross validation has facilitated the choice.

| Type of Indice | Value |
|---|---|
| Mean Error | 0.205614 |
| Mean square error | 10.45415 |

The green straight lines in the Figure 3 and 4 is the regression straight lines of NOx values estimated

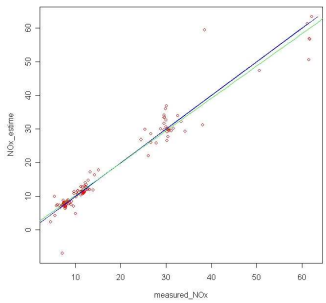



Figure 3: The scatter plot of the observed and predicted values of NOx, using the Kriging method (first model).

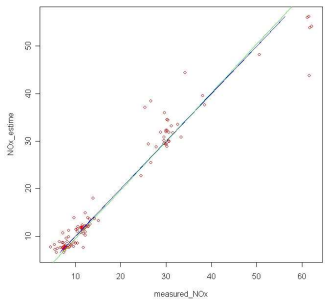



Figure 4: The scatter plot of the observed and predicted values of NOx, using the Kriging method (second model).

by the model, on the values of NOx observed. We notice that this straight lines coincide with the Black straight lines given by equation $y = x$, which is why the estimate obtained is effective

Both experimental variogram calculated in the framework of these two approaches are almost similar. We notice that the two sructures spatial show a strong dependence.

In both cases the estimations and the results given by cross-validation are good. On the other hand, it is clear that, for some experiments, the first one is the best, and, for other experiments, it is the second one which gives the best results.

These results impel us to adopt, in future work, a combination of these models in order to optimize the estimation of NOx emissions.

## 5 CONCLUSIONS

This paper describes a pollutant emissions simulator of compression ignition engine. The effort has been put into building a model based on the kriging method. The resulting model can predict engine pollutant emissions and can be used to predict the engine performance and noise, it is easy to generalize for various diesel engine configurations. This model is also suitable for real time simulations. The predictions obtained by this simulator are satisfactory compared to the results obtained by using of a physical model given by S.castric et al (Castric, 2007). Our future aim is to estimate the engine performance by using the proposed model. This latter will be adopted in order to make the multi-objective optimization, using the stochastic methods.

## REFERENCES

Castric, S. (2007). Readjusting methods for models and application for diesel emissions. In *PhD thesis, Universit Technologique de Compiègne*.

Davis.J.C (1986). *Statistics and Data Analysis in Geology*. John Wiley and Sons, New York, 2nd edition.

Heywood, J. (1988). *Internal combustion engine fundamentals*. Mac Graw-Hill, London.

S. Baillargeon, J. P. (2004). interpolation statistique multivariable de données de prcipitations dans un cadre de modlisation hydrologique. In *Colloque Gomatique 2004: un choix stratègique, Montrèal*.

S. Castric, V. T. (2007). A diesel engine combustion model for tuning process and a calibration method. In *IMSM07 The Third Int. Con. on AVCS'07, Buenos Aires, Argentine*.

# OFF-LINE ROBUSTIFICATION OF PREDICTIVE CONTROL FOR UNCERTAIN SYSTEMS
## *A Sub-optimal Tractable Solution*

Cristina Stoica, Pedro Rodríguez-Ayerbe and Didier Dumur
*Department of Automatic Control, Supélec, 3 rue Joliot Curie, F91192 Gif-sur-Yvette, France*
*cristina.stoica@supelec.fr, pedro.rodriguez@supelec.fr, didier.dumur@supelec.fr*

Keywords: Model predictive control, Multivariable systems, Polytopic uncertainties, Robust control, LMIs, BMIs.

Abstract: An off-line technique enabling to robustify an initial Model Predictive Control (MPC) for multivariable systems via the convex optimization of a Youla parameter is presented. Firstly, a multivariable predictive controller is designed for a nominal system and then robustified towards unstructured uncertainties, while guaranteeing stability properties over a specified polytopic domain of uncertainties. This condition leads to verify a Bilinear Matrix Inequality (BMI) for each vertex of the polytopic domain. This BMI can be mathematically relaxed to semi-definite programming (SDP) using a Sum of Squares (SOS) strategy, with a significant increase of the number of scalar decision variables. To overcome this inconvenient, an alternative tractable sub-optimal solution for the BMI is proposed, based on the elaboration of a stable solution obtained by minimization of the complementary sensitivity function.

## 1 INTRODUCTION

During the latest years, the robustness aspect of Model Predictive Control (MPC) has been considered both within online strategies (Kothare *et al.*, 1996; Goulart and Kerrigan, 2007; Camacho and Bordons, 2004) and off-line approaches (Wan and Kothare, 2003; Rossiter, 2003; Rodríguez and Dumur, 2005). Mixed methods computing off-line a set of controllers have been developed, leaving on-line only the selection of the current controller (Olaru and Dumur, 2004; Lee and Kouvaritakis, 2006).

This paper presents an off-line robustification procedure for model predictive control applied to multivariable (possibly non-square) uncertain systems. It considers both unstructured and polytopic uncertainties. Firstly, a predictive controller for a nominal system is designed. Secondly, the robustification problem under unstructured uncertainties is considered. This leads to a convex optimization of a multivariable Youla parameter solved with Linear Matrix Inequalities (LMIs) techniques, as described in (Stoica *et al.*, 2007). Thirdly, the robust stability of the controlled system towards system polytopic uncertainties is considered. Since the polytopic domain is chosen as a convex polytope, this implies checking the stability only for the vertices of the polytope (Kothare *et al.*, 1996). This condition leads to satisfy a Bilinear Matrix Inequality (BMI) for all vertices of the polytopic domain. This problem can be transformed into semi-definite programming (SDP) using Sum of Squares (SOS) relaxations described in (Scherer and Hol, 2006), with a significant increase of the number of scalar decision variables. To avoid this increase of the computing time, this paper proposes a sub-optimal tractable solution based on the minimization of the complementary sensitivity function which permits to enlarge the stability domain. A feasible solution for each vertex can be found, the stability conditions for all the vertices of the polytopic domain being then explicitly integrated.

The most interesting result is that this robustification technique permits to guarantee the stability property on the entire polytopic uncertain domain, even if the initial MPC controller may be unstable for some regions of the polytopic domain.

This paper is organized as follows. The main steps leading to a MIMO MPC and the related class of stabilizing controllers are presented in Section 2. The robustification procedure under unstructured and polytopic uncertainties is detailed in Section 3. Finally, some concluding remarks are given in Section 4.

## 2 CLASS OF STABILIZING MPC

This section briefly presents the main steps leading to an initial stabilizing multivariable MPC in state-space formalism and the class of stabilizing controllers obtained via the Youla parameter. More details can be found in (Stoica *et al.*, 2007). Let us consider a discrete time MIMO LTI system with $m$ inputs and $p$ outputs, characterized by the 4-uplet $(\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{0})$ of the state-space representation.

In order to cancel the steady-state errors, an integral action on the control vector is added, leading to an extended state-space description:

$$\begin{cases}\mathbf{x}_e(k+1)=\mathbf{A}_e\,\mathbf{x}_e(k)+\mathbf{B}_e\,\Delta\mathbf{u}(k)\\ \mathbf{y}(k)=\mathbf{C}_e\,\mathbf{x}_e(k)\end{cases} \tag{1}$$

Minimizing the quadratic objective function (2) gives the expression of the control signal. The following notations are used: $\mathbf{y}_r$ - the setpoint; $\widetilde{\mathbf{Q}}_J$, $\widetilde{\mathbf{R}}_J$ - the weighting matrices. The future control increments $\Delta\mathbf{u}(k+i)$ are supposed to be 0 for $i \geq N_u$. The same output prediction horizons ($N_1$, $N_2$) and the same control horizon $N_u$ are applied for all input/output transfers.

$$J=\sum_{i=N_1}^{N_2}\left\|\hat{\mathbf{y}}(k+i)-\mathbf{y}_r(k+i)\right\|^2_{\widetilde{\mathbf{Q}}_J(i)}+\sum_{i=0}^{N_u-1}\left\|\Delta\mathbf{u}(k+i)\right\|^2_{\widetilde{\mathbf{R}}_J(i)} \tag{2}$$

The predicted output $\hat{\mathbf{y}}(k)$ is derived from:

$$\hat{\mathbf{y}}(k+i)=\mathbf{C}\,\mathbf{A}^i\hat{\mathbf{x}}(k)+\sum_{j=0}^{i-1}\mathbf{C}\,\mathbf{A}^{i-j-1}\mathbf{B}\mathbf{u}(k+j) \tag{3}$$

with $\hat{\mathbf{x}}(k)$ obtained from the following observer:

$$\hat{\mathbf{x}}_e(k+1)=\mathbf{A}_e\,\hat{\mathbf{x}}_e(k)+\mathbf{B}_e\,\Delta\mathbf{u}(k)+\mathbf{K}[\mathbf{y}(k)-\mathbf{C}_e\,\hat{\mathbf{x}}_e(k)] \tag{4}$$

An analytical minimization of (3) rewritten in a matrix form, as described in (Maciejowski, 2001), leads to the following control signal (Fig. 1):

$$\Delta\mathbf{u}(k)=\mathbf{F}_w\mathbf{w}(k)-\mathbf{L}\,\hat{\mathbf{x}}_e(k) \tag{5}$$

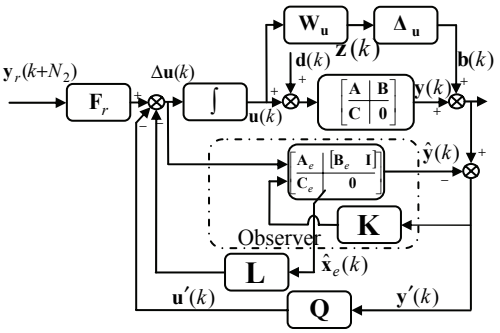



Figure 1: Robustified MIMO MPC via **Q** parametrization.

The structure of the control gain matrix $\mathbf{L}=[\mathbf{L}_1\ \mathbf{L}_2]$ and the setpoint pre-filter $\mathbf{F}_w$ are the same as in (Stoica *et al.*, 2007). The expression (5) provides an initial stabilizing controller. A possible way leading to the class of all stabilizing controllers is to use the Youla-Kučera parameter coupled with this control law. It is well known from the literature (Boyd and Barratt, 1991; Maciejowski, 1989) that any stabilizing controller can be represented by a state-space feedback controller coupled with an observer and a Youla (also called **Q**) parameter.

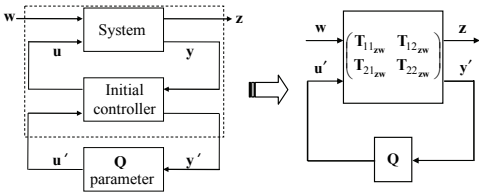



Figure 2: Class of stabilizing controllers with Q parameter.

The first step is to add supplementary inputs $\mathbf{u}'$ and outputs $\mathbf{y}'$ with a zero transfer between them ($\mathbf{T}_{22_{\mathbf{zw}}}=0$ in Fig. 2), which permits the connection of the **Q** parameter between $\mathbf{y}'$ and $\mathbf{u}'$ without restricting the closed-loop stability. As a result, the closed-loop function between **w** and **z** is linearly parametrized by the **Q** parameter, allowing convex specification (Boyd and Barratt, 1991):

$$\mathbf{T}_{\mathbf{zw}}=\mathbf{T}_{11\,\mathbf{zw}}+\mathbf{T}_{12\,\mathbf{zw}}\,\mathbf{Q}\,\mathbf{T}_{21\,\mathbf{zw}} \tag{6}$$

where $\mathbf{T}_{11_{\mathbf{zw}}},\mathbf{T}_{12_{\mathbf{zw}}},\mathbf{T}_{21_{\mathbf{zw}}}$ depends on the considered input/output ($\mathbf{w}/\mathbf{z}$) transfer.

# 3 ROBUSTNESS VIA THE YOULA PARAMETRIZATION

A procedure enhancing robustness of the previous multivariable MPC in terms of the Youla parameter is presented in the particular case of the maximization of the robust stability under additive unstructured uncertainties, while guaranteeing stability properties over a specified polytopic domain of uncertainties. It will be shown that the global robustification problem is a necessary trade-off between both robustification aspects.

## 3.1 Robust Stability under Unstructured Uncertainties

Along with the small gain theorem (Maciejowski, 1989; Zhou *et al.*, 1996), a necessary and sufficient condition for the robust stability under unstructured uncertainties $\Delta_u$ (Fig 3) is formulated as the following $H_\infty$ norm minimization:

$$\min_{\mathbf{Q}\in\Re H_\infty} \left\|\mathbf{T_{zw}}\right\|_\infty \tag{7}$$

where $\Re H_\infty$ is the space of stable transfers and $\mathbf{T_{zw}}$ also contains the weighting factors.

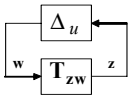



Figure 3: Unstructured uncertainty.

The minimization (7) may be more specifically formulated using the following theorem.

**Theorem** (Clement and Duc, 2000; Boyd *et al.*, 1994): A discrete time system given by the state-space representation $(\mathbf{A}_{cl},\mathbf{B}_{cl},\mathbf{C}_{cl},\mathbf{D}_{cl})$ is stable and admits a $H_\infty$ norm lower than $\gamma$ if and only if:

$$\exists \mathbf{X}_1 = \mathbf{X}_1^{\mathrm{T}} \succ 0 / \begin{bmatrix} -\mathbf{X}_1^{-1} & \mathbf{A}_{cl} & \mathbf{B}_{cl} & \mathbf{0} \\ \mathbf{A}_{cl}^{\mathrm{T}} & -\mathbf{X}_1 & \mathbf{0} & \mathbf{C}_{cl}^{\mathrm{T}} \\ \mathbf{B}_{cl}^{\mathrm{T}} & \mathbf{0} & -\gamma\mathbf{I} & \mathbf{D}_{cl}^{\mathrm{T}} \\ \mathbf{0} & \mathbf{C}_{cl} & \mathbf{D}_{cl} & -\gamma\mathbf{I} \end{bmatrix} \prec 0 \tag{8}$$

where the notation "$\succ 0$"/"$\prec 0$" refers to a strictly positive/negative definite matrix. There exist appropriate techniques to transform the expression (8) into a LMI (Clement and Duc, 2000; Scherer, 2000). The decision variables should be $\mathbf{X}_1$, $\gamma$ and the **Q** parameter included in the closed-loop matrices (Stoica *et al.*, 2007). As a result, the optimization problem is formulated as the minimization of $\gamma$ subject to this first LMI constraint:

$$\min_{LMI_0} \gamma \tag{9}$$

To restrict the search of the **Q** parameter which initially varies in the infinite-dimensional space $\Re H_\infty$, a sub-optimal solution is to consider for each input/output pairs $(i,j)$ a finite-dimensional subspace generated by an orthonormal base of discrete stable transfer functions (such as a polynomial or FIR filter). This MIMO Youla parameter can be obtained in the state-space formalism using a fixed pair $(\mathbf{A}_Q,\mathbf{B}_Q)$ and searching only for the variable pair $(\mathbf{C}_Q,\mathbf{D}_Q)$.

## 3.2 Robust Stability under Polytopic Uncertainties

The main result is the robustification procedure under polytopic uncertainties. Consider the following time-varying system, as a generalization of the polytopic system (Kothare *et al.*, 1996):

$$\begin{cases} \mathbf{x}(k+1) = \mathbf{A}(k)\mathbf{x}(k) + \mathbf{B}(k)\mathbf{u}(k) \\ \mathbf{y}(k) = \mathbf{C}(k)\mathbf{x}(k) \end{cases} \tag{10}$$

where $\left[\mathbf{A}(k)\ \mathbf{B}(k)\ \mathbf{C}(k)\right] \in \boldsymbol{\Omega}$ and the polytope $\boldsymbol{\Omega}$ (Fig. 4) represents the convex hull *Co* defined by the $l$ vertices $\left[\mathbf{A}_i\ \ \mathbf{B}_i\ \ \mathbf{C}_i\right]$.

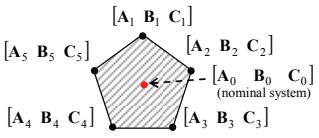



Figure 4: Polytopic uncertainty representation ($l = 5$).

As $\boldsymbol{\Omega}$ is a polytope (convex set), guaranteeing the stability of (10) on the entire space $\boldsymbol{\Omega}$ means to guarantee the stability for all the vertices of the polytope (Kothare *et al.*, 1996). This is equivalent to satisfy the following condition (Boyd *et al.*, 1994) for each vertex $i = \overline{1,l}$ of the domain $\boldsymbol{\Omega}$:

$$\begin{bmatrix} -\mathbf{X}_2 & \mathbf{X}_2\mathbf{A}_{cl,i} \\ \mathbf{A}_{cl,i}^{\mathrm{T}}\mathbf{X}_2 & -\mathbf{X}_2 \end{bmatrix} \prec 0,\ \ \mathbf{X}_2 = \mathbf{X}_2^{\mathrm{T}} \succ 0 \tag{11}$$

This expression is bilinear in its decision variables $\mathbf{X}_2$ and the **Q** parameter included in $\mathbf{A}_{cl,i}$. The global robustification problem towards both unstructured and polytopic uncertainties is achieved

by minimizing $\gamma$ subject to the constraints $LMI_0$ and $BMI_i$ (11):

$$\min_{LMI_0, BMI_i, i=\overline{1,l}} \gamma \tag{12}$$

But this is a difficult problem since it involves BMI expressions, in addition containing decision variables (the **Q** parameter) jointly with a LMI. The challenge is to try to find a sub-optimal solution.

A first mathematical approach based on Sum of Squares (SOS) for relaxing the BMIs (12) is developed in the literature by (Scherer and Hol, 2006). But this relaxation technique leads to a huge number of scalar decision variables (that Matlab™ cannot deal with it for the moment) due to the size of SOS matrices. Hence it cannot be used within the presented robustification procedure.

For this reason, a second sub-optimal tractable solution (in three steps) of solving these BMIs is proposed. Firstly, in order to enlarge the polytopic domain around the nominal system, the minimization of the complementary sensitivity function is added to (9). This is equivalent to add the minimization of the transfer between **b** and **y** (Fig. 1) to (9). This minimization is then transformed into a LMI added to the first one (9):

$$\min_{LMI_0, LMI_{CS}} c_1\gamma + c_2\gamma_{CS} \tag{13}$$

choosing appropriate coefficients $c_1, c_2$. Solving the optimization problem (13) leads to a **Q** parameter that will be used in the second step of the robustification procedure. In fact, the minimization (13) is recomputed until the resulting stability domain includes at least the polytopic domain of uncertainties, by selecting appropriate weightings $c_1, c_2$. The expression (13) offers the possibility to increase the stability domain, but does not offer any information about the limits of this domain. To explicitly include the considered polytopic domain, the second and third steps must be followed.

In order to find a sub-optimal solution of (11), the second step is to search $\mathbf{X}_2$ using the **Q** parameter obtained with (13). This can be achieved for instance by minimizing the trace of $\mathbf{X}_2$ subject to the $LMI_i$ ($i = \overline{1,l}$) derived from the BMIs (11), which permits to choose $\mathbf{X}_2$ in order to enlarge the stability domain:

$$\min_{LMI_i,\ i=\overline{1,l}} tr(\mathbf{X}_2) \tag{14}$$

Thirdly, the value obtained for $\mathbf{X}_{2,i}$ is used in the final step of the optimization problem which decision variables are $\mathbf{X}_1$, $\gamma$ and the **Q** parameter included in the closed-loop matrices from $LMI_0$ and $LMI_i$:

$$\min_{LMI_0, LMI_i,\ i=\overline{1,l}} \gamma \tag{15}$$

where $LMI_i$ are the relaxations of the BMIs (11) for the vertices $\mathbf{A}_i$, while fixing the variable $\mathbf{X}_2$. The optimization (15) gives a Youla parameter that will guarantee the stability of the controlled system for all the vertices of the polytopic domain.

# 4 CONCLUSIONS

This paper has proposed an off-line methodology which improves the robustness of an initial stabilizing predictive controller via the convex optimization of the Youla parameter. This procedure deals with the stability robustness aspect of the nominal system towards unstructured uncertainties (solved with LMI tools), while guaranteeing the stability under a considered polytopic uncertain domain (leading to BMIs). In order to find a sub-optimal solution for these BMIs, a new method presenting a sub-optimal technique of solving this non-convex problem is proposed: one matrix variable is fixed using the minimization of the complementary sensitivity function, while looking for the other matrix variable. This provides computationally tractable solutions.

The main advantage of this robustification technique under polytopic uncertainties is that guaranteeing the BMI stability condition robustly stabilizes the controlled system for the entire polytopic domain, even if the system coupled with the initial predictive controller is unstable in some points of the polytopic domain. This offers a possible way of increasing the polytopic domain for which the stability is guaranteed.

# REFERENCES

Boyd, S., Barratt, C., 1991. *Linear controller design. Limits of performance*, Prentice Hall.

Boyd, S., Ghaoui, L.El., Feron, E., Balakrishnan, V., 1994. *Linear matrix inequalities in system and control theory*, SIAM Publications, Philadelphia.

Camacho, E.F., Bordons, C., 2004. *Model predictive control,* Springer-Verlag. London, 2nd edition.

Clement, B., Duc, G., 2000. A multiobjective control via Youla parameterization and LMI optimization:

application to a flexible arm, *IFAC Symposium on Robust Control and Design*, Prague.

Goulart, P.J., Kerrigan E.C., 2007. Output feedback receding horizon control of constrained systems. *International Journal of Control*, *80(1)*, pp. 8-20.

Kothare, M.V., Balakrishnan, V., Morari, M., 1996. Robust constrained model predictive control using linear matrix inequalities. *Automatica*, *32(10)*, pp. 1361-1379.

Lee, Y.I., Kouvaritakis, B., 2006. Constrained robust model predictive control based on periodic invariance. *Automatica, 42*, pp. 2175-2181.

Maciejowski, J.M., 1989. *Multivariable feedback design*, Addison-Wesley Publishing Company, Wokingham.

Maciejowski, J.M., 2001. *Predictive control with constrains*, Prentice Hall.

Olaru, S., Dumur, D., 2004. A parameterized polyhedra approach for explicit constrained predictive control. *43rd IEEE CDC*, The Bahamas.

Rodríguez, P., Dumur, D., 2005. Generalized predictive control robustification under frequency and time-domain constraints. *IEEE Transactions on Control Systems Technology, 13(4)*, pp. 577-587.

Rossiter, J.A., 2003. *Model-based predictive Control. A practical approach.* CRC Press LLC.

Scherer, C.W., 2000. An efficient solution to multi-objective control problem with LMI objectives. *Systems and Control Letters, 40*, pp 43-57.

Scherer, C.W., Hol C.W.J., 2006. Matrix sum-of-square relaxations for robust semi-definite programs. *Math. Program.*, *Ser. B 107*, pp 189-211.

Stoica, C., Rodríguez-Ayerbe, P., Dumur, D., 2007. Improved robustness of multivariable model predictive control under model uncertainties. *4th ICINCO*, Angers.

Wan, Z., Kothare, M., 2003. An efficient off-line formulation of robust model predictive control using linear matrix inequalities. *Automatica*, *39(5)*, pp 837-846.

Zhou, K., Doyle, J.C., Glover, K. 1996. *Robust and optimal control.* Prentice Hall.

# REAL-TIME SYSTEMS SAFETY CONTROL CONSIDERING HUMAN MACHINE INTERFACE

José Machado and Eurico Seabra
*Mechanical Engineering Department, University of Minho, Campus of Azurém, 4800-058 Guimarães, Portugal*
*{jmachado, eseabra}@dem.uminho.pt*

Keywords: Real-Time Systems, Safety Control, Human Machine Interface, Dependable Systems.

Abstract: In this paper it is presented the analysis of real-time industrial controllers when it is taken into account human behavior in the use of fully automated industrial systems. It is intended to develop safe controllers for these systems and make them robust against inappropriate utilizations by human operators. For the attainment of our goals it is used a case study, where, based on a IEC 60848 specification, is deduced the controller program. Further, it is elaborated the controller model, the Plant model and the Human Machine Interface Model of the automated system. The obtained results are generalized for other similar systems with the presented case study.

## 1 INTRODUCTION

Since the early eighties, the influence of the human role and of the degree of the human implication in the human-machine global performance (production, safety,…) has been studied. Tom Sheridan defined (Sheridan, 1984) the well-known degrees of automation and their consequences. These defined three degrees of automation and their consequences are:

- In fully manual controlled systems, safety totally depends on the human controller reliability;
- An intermediate, state allows a task sharing between the human operators and the automated controlled systems; and,
- Fully automated systems reject the human operator out of the control and that can produce a lack of vigilance, a loss of skill and can prevent him to assume all the responsibility on the system. Therefore, the system safety is almost totally linked to the technical reliability.

In the study presented on this paper we will focus on the third point related with the fully automated systems. In the industrial controllers analysis, it will be used simulation and formal verification techniques to increase, together, the safety of industrial controllers.

Among the several available techniques for the industrial controllers analysis, Simulation (Baresi *et al.* 2000, Baresi *et al.* 2002) and Formal Verification (Moon 1994, Roussel and Denis 2002), can be distinguished due to their utility. In the research works on industrial controllers' analysis, these two techniques are rarely used simultaneously. In our work, here presented, these two techniques are used together and it is shown that exist some limitations in the use of Simulation when compared with Formal Verification. These limitations are demonstrated in a context of studying the Human Machine Interface (HMI) of real time industrial systems. Some results that seam to be correct, using Simulation, may not be correct when using Formal Verification. This paper is focused in giving an overview in the limitations of Simulation when compared with Formal Verification in a context of the HMI.

To accomplish our 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 a general presentation of the case study involving a system with two tanks filled and emptied by the control of some on-off valves. Further, it is presented the methodology to obtain the controller program deduced from an IEC 60848 SFC specification of the system's desired behaviour. Sections 3 and 4 are, respectively, devoted to the modelling the plant and HMI. In section 5 are presented the system behaviour analysis results and

finally, in section 6 are presented some conclusions and future work.

## 2 CASE STUDY

The case study, composed by two tanks and filling and emptying on-off valves, is presented in figure 1. Tank1 is filled by opening valve V1. When the level of the tank1 becomes high, the valve V1 is closed. After a waiting time of ten time units, valve V2 is opened and the fluid flows from tank1 into tank2.

When tank1 is empty, valve2 is closed and, after a waiting time of fifteen time units, valve3 is opened and the fluid flows out of tank2. Finally when tank3 is empty, valve V3 is closed. In this work we consider that one time unit is equal to one second.

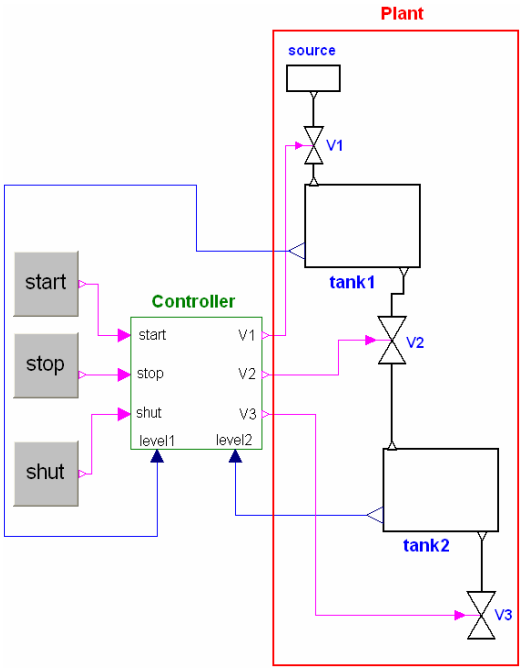



Figure 1: Evaporator system: Closed-loop system composed by controller and plant and start, stop and shut buttons to interact with human behaviour.

Three buttons can influence the above normal operation: *start*, *stop* and *shut*. In order to guarantee the desired functioning of the system it is necessary to simulate the following desired behaviors, traduced by three system behavior properties:

- Property 1 (**P1**): In the beginning, when the *start* button is pressed the system must start, immediately, filling the tank 1.
- Property 2 (**P2**): Button "shut" is used to shutdown the process. When the *shut* button is pressed the system controller must reach the initial state or the system must begin emptying immediately the two tanks, in simultaneous.
- Property 3 (**P3**): When the *stop* button is pressed the system must stand in its actual situation.

### 2.1 IEC 60848 Controller Specification

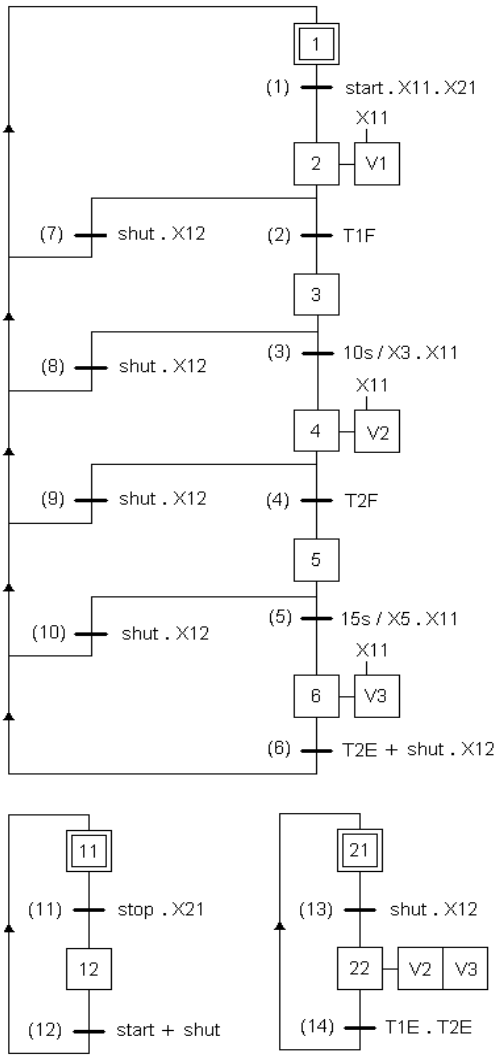



Figure 2: SFC specification of the controller.

As we use the Simulation and Formal Verification, using different tools and intending to conciliate the obtained results, we adopted a controller specification that is the same for the basis of the controller program in the two analysis techniques.

Thus, the controller specification was developed in IEC 60848 SFC because it can be used as the basis for the development of the Programmable

Logic Controller program (PLC), to be verified with UPPAAL based on timed automata (Alur and Dill, 1990), and also it is the basis for the controller program to be used by StateGraphs Modelica library (Otter *et al.* 2005).

Table 1: Input/Output variables of the controller.

| Inputs | Outputs |
|---|---|
| Start – system start<br>stop – system stop<br>shut – system shutdown<br>T1E – tank1 empty<br>T1F – tank1 full<br>T2E – tank1 empty<br>T2F – tank1 full | V1 – open valve1<br>V2 – open valve2<br>V3 – open valve3 |

The input and output variables of the controller model are summarized on Table 1; minimum and maximum level sensors of the tanks and the human-machine interface buttons (*start*, *stop* and *shut*) are controller program inputs and the on-off valves (V1, V2 and V3) are controller program outputs.

In order to guarantee the desired behaviour for the described system, a IEC 60848 SFC specification is presented in Figure 2. As IEC 60848 SFC is a specification language (and not a programming one), it is necessary to translate the SFC specification, first to a StateGraph program, presented in (Seabra and Machado, 2007) and, second, to translate it into a program written in a PLC programming language (in this case it will be used the ladder language). This translation is done using a methodology, having as base the specification algebraic representation and considering also the controller program behavior presented at (Machado, 2006).

# 3 MODEL OF THE PLANT

In the plant modeling, first, the plant is modeled with the Modelica programming language (Elmqvist and Mattson, 1997) and simulated with the Dymola software and, second, it is modeled by timed automata to be used as input of the UPPAAL software (David *et al.* 2003). The delays obtained, in the simulation with the Dymola software, are used to create the timed automata that are used on Formal Verification with the UPPAAL tool.

## 3.1 Plant Modelling for Simulation Purposes

All the system was modeled. The tank1 model is presented on this sub-chapter. Lets consider the case of the tank 1 we have, for the Modelica programming language the model presented in Figure 3.

```
model Tank1
  Modelica.Blocks.Interfaces.RealOutput levelSensor;
  Modelica.StateGraph.Examples.Utilities.inflow inflow1;
  Modelica.StateGraph.Examples.Utilities.outflow outflow1;
  Real level "Tank level in % of max height";
  parameter Real A=1 "ground area of tank in m²";
  parameter Real a=0.2 "area of drain hole in m²";
  parameter Real hmax=1 "max height of tank in m";
  constant Real g=Modelica.Constants.g_n;
  Modelica.StateGraph.Examples.Utilities.inflow inflow2;
equation
  der(level) = (inflow1.Fi + inflow2.Fi - outflow1.Fo)/(hmax*A);
  if outflow1.open then
    outflow1.Fo = sqrt(2*g*hmax*level)*a;
  else
    outflow1.Fo = 0;
  end if;
  levelSensor = level;

end Tank;

connector Modelica.Blocks.Interfaces.RealOutput =
                    output RealSignal "'output Real' as connector";

connector Modelica.Blocks.Interfaces.RealSignal
  "Real port (both input/output possible)"
  replaceable type SignalType = Real;

  extends SignalType;

end RealSignal;

connector Modelica.StateGraph.Examples.Utilities.inflow

    import Units = Modelica.SIunits;

  Units.VolumeFlowRate Fi "inflow";
end inflow;

connector Modelica.StateGraph.Examples.Utilities.outflow

    import Units = Modelica.SIunits;

  Units.VolumeFlowRate Fo "outflow";
  Boolean open "valve open";
end outflow;
```

Figure 3: Modelica code for tank1 model.

The other physical parts of the system were modeled (Seabra and Machado, 2007) but not presented in this paper because is not part of the goals of this paper.

## 3.2 Plant Modelling for Formal Verification Purposes

For modeling the plant with formal verification purposes there are considered the following modules for the plant modeling: Tank1 and Tank2.

Model of tank1:

The obtained delays on simulation were used on formal verification with UPPAAL. The corresponding model of the tank developed in UPPAAL for formal verification purposes is presented in Figure 4.

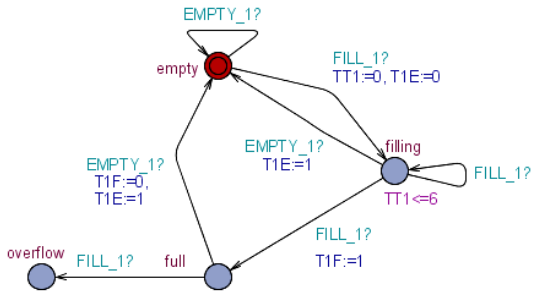



Figure 4: UPPAAL model of tank1.

We consider four states: *empty* models that tank1 is empty; *filling* models that the liquid is entering in tank1; *full* models that tank1 is full; state *overflow* is also considered, this is a possible state for the tank, but describes an undesired behaviour. In this model, it is also considered that the tank1 is emptied in a very short time, when compared with the filling time. We have considered this time null. It is for that reason that the model goes from the *full* state directly to the *empty* state, without an intermediate state. The Boolean variables T1E and T1F are associated with *tank1.empty* and *tank1.full,* respectively. These variables represent the level sensors' signals sent by the sensors from the plant to the controller. The maximum time for filling tank1 is six time units.

Model of tank2:

The model of tank2 is presented in figure 5 and the reasoning followed to obtain this model was the same as presented before for obtaining the tank1 model. As empting tank1 is considered to take a short (null) time, the filling of the tank2 is done in the same conditions, since the liquid is transferred from tank1 to tank2.

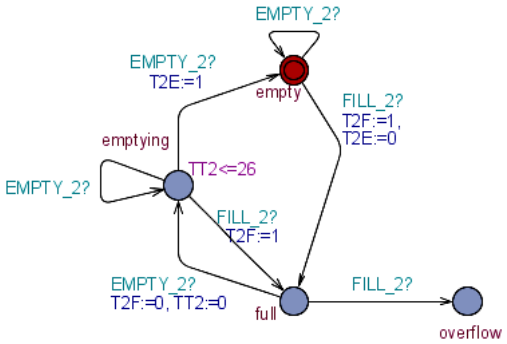



Figure 5: UPPAAL model of tank2.

Four states are considered: *empty*, *full*, *emptying* and *overflow* which is a possible state for the tank, but describes an undesired behavior. The variables T2E and T2F have the same behavior on the tank2 model as the T1E and T1F described above on the tank1 model. Empty tank2 takes, at maximum, twenty-six time units.

# 4 MODELLING THE HUMAN MACHINE INTERFACE

In this chapter are modelled the three buttons: *start*, *stop* and *shut*. The models for these elements of the (HMI) are presented in figures 6, 7 and 8.

For each HMI button the considered behaviours are that each one can be in the state *off* or *on*. They can change of state at any time, according the human behaviour.

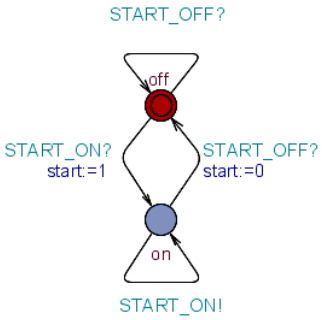



Figure 6: Model of the *start* button.

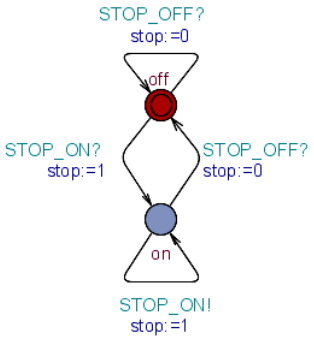



Figure 7: Model of the *stop* button.

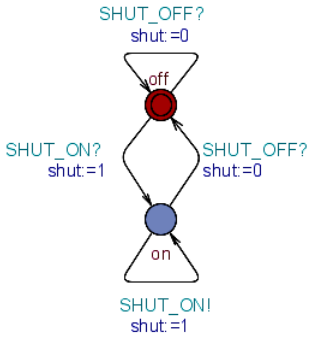



Figure 8: Model of the *shut* button.

In the evolutions of the controller model it is considered that it will be implemented into a PLC with a scan cycle composed by three distinct phases: *Controller Inputs Reading* (CIR), *Controller Computing* (CCO) and *Controller Outputs Updating* (COU).

The evolution of the controller model takes into account the state changing of the HMI buttons at each moment that it is in the CIR state. Any changing of the HMI buttons state during the evolution of the controller behaviour is not detected by the controller model (as it is in the real behaviour of the controller). Taking into account the characteristics of the controller behaviour, the characteristics of the plant behaviour and the characteristics of the HMI behaviour, the properties must be proved in the end of the evolution of the controller model, after the *Controller Outputs Updating* states.

# 5 SYSTEM BEHAVIOUR ANALYSIS RESULTS

In the analysis of the system behaviour there were used the two indicated techniques: the Simulation and the Formal Verification.

It is pointed out that the system behaviors that we intend to analyze, in this paper, are directly related with possible human behaviors (correct or incorrect) in the use of the automated system. There are allowed, in the HMI models all possible human behaviors with the three considered buttons.

The work presented here is a small part of a larger developed work in the proof of properties in real-time systems. The developed work in the context of the controller behavior properties and the global system behavior properties were presented, respectively, in (Machado et al. 2007-a) and (Machado et al. 2007-b).

The properties to prove, related with HMI, are:

- Property 1 (**P1**): In the beginning, when the start button is pressed the system must start, immediately, filling the tank 1.
- Property 2 (**P2**): Button "shut" is used to shutdown the process. When the shut button is pressed the system controller must reach the initial state or the system must begin emptying immediately the two tanks, in simultaneous.
- Property 3 (**P3**): When the stop button is pressed the system must stand in its actual situation.

## 5.1 Simulation Results

Considering Simulation all properties are true.

As in this paper it is intended to show the advantages of Formal Verification related with Simulation, we will focus on Formal Verification results discussion.

## 5.2 Formal Verification Results

Before presenting the formal verification results the properties must be formalized using the UPPAAL syntax, and, for that, we need a small part of TCTL formalism (Alur *et al.* 1993 ). In the formulas below which are all (possibly timed) invariants, A is the universal quantifier on paths: for *any path*…, and [ ] means *always*… The combination A [ ] means *for all states in the future*…

There are considered, for the properties formalization, the input and output variables of the controller, the step variables of the controller SFC program and the state of the controller model, where are verified the properties according some rules defined in (Machado, 2006). The considered state of the controller model for the properties verification is the *Controller Outputs Updating* state, COU.

For the properties formalization we have:

- **P1**: A[] !(COU && X1 && start)
- **P2**: A[] !((COU && !X22 && shut) || (COU && !X1 && shut))
- **P3**: A[] !(COU && X11 && stop)

After the formal verification tasks, the obtained results are that all the properties are false.

## 5.3 Discussion of the Obtained Results

All the results, that are false in Formal Verification analysis, are related with the controller behaviour (Cyclic scan monitor of the PLC).

Indeed, in Simulation, this detail is not taken into account but, in Formal Verification (because it is an exhaustive technique!), these undesired behaviours are detected and we can show that, with this technique, the obtained results are exhaustive and precise.

Detailing the results obtained for the Property 1 we can interpret the obtained trace with the following sequence of human operator actions (see figure 2):

In the initial situation, of the system, if the human operator does not press the *start* button (as expected!) but presses the *stop* button, the step 12 is activated;

If, after that, the human operator presses the *shut* button the step 22 is activated too;

Further, if the human operator presses the *start* button the system does not start its normal behaviour because the variable step X21 is not activated and the step 1 remains active at least during a PLC internal cycle.

For all the other properties the obtained results (false) may be explained in the same way, when analyzing the respective traces.

To solve this problem, there are many possibilities. The simpler one seems to consider actuation priorities for the three buttons considered. These priorities must be included on the controller program specification and, consecutively, in the controller program implementation.

# 6 CONCLUSIONS AND FUTURE WORK

With our study it has been possible to show that some problems can occur if the development of safe industrial controllers, for fully automated systems, are not developed taking into account some possible incorrect behaviours of human operators.

These possible undesired system behaviours can be detected, only, if it is used the Formal Verification technique; the Simulation technique is not sufficient.

The fully automated systems safety is almost totally linked to the technical reliability of the system and it must be guaranteed that some incorrect possible behaviours of the human operators do not compromise these systems' safety and dependability.

# ACKNOWLEDGEMENTS

This research project is carried out in the context of the SCAPS Project supported by FCT, the Portuguese Foundation for Science and Technology, and FEDER, the European regional development fund, under contract POCI/EME/61425/2004 that deals with safety control of automated production systems.

# REFERENCES

Alur R., Dill D. L., 1990. Automata for Modeling Real-Time Systems. Proceedings of the *17th Int. Coll. Automata, Languages, and Programming (ICALP'90)*, Warwick University, England, July 1990, Vol. 443, Lecture Notes in Computer Science, Springer.

Alur R., Courcoubetis C., Dill D. L, 1993. Model-Checking in Dense Real-Time. *Information and Computation, vol. 104, n_ 1, p. 2-34*.

Baresi L., Mauri M., Monti A., Pezzè M., 2000. PLCTOOLS: Design, Formal Validation, and Code Generation for Programmable Controllers. *Special Session at IEEE Conference on Systems, Man, and Cybernetics*. Nashville USA.

Baresi L., Mauri M., Pezzè M., 2002. *PLCTools: Graph Transformation Meets PLC Design*. Electronic Notes in Theoretical Computer Science 72 No. 2.

David A., Behrmann G., Larsen K. G., Yi W., 2003. *A Tool Architecture for the Next Generation of UPPAAL*. Technical Report n. 2003-011, Department of Information Technology, Uppsala University, Feb. 20 pages.

Elmqvist E., Mattson S., 1997. An Introduction to the Physical Modelling Language Modelica. *Proceedings of the 9th European Simulation Symposium, ESS'97*. Passau, Germany.

Machado J., 2006. Influence de la Prise en Compte d'un Modèle de Processus en Vérification Formelle des Systèmes à Evénements Discrets. *PhD Thesis in cooperation between the University of Minho and École Normale Supérieure de Cachan*; School of Engineering, University of Minho, June.

Machado J., Seabra E., Campos J., Soares F., Leão C., Silva J., 2007-a. Simulation and Forml Verification of Industrial Systems Controllers. Proceedings of *19th Edition of the International Congress of Mechanical Engineering (COBEM'2007)*, Brazilia, Brazil, 5-9th November.

Machado J., Seabra E., Soares F., Campos J., 2007-b. A new Plant Modelling Approach for Formal Verification Purposes. Proceedings of the *11th IFAC/IFORS/IMACS/ IFIP Symposium on Large Scale Systems: Theory and Applications*. Gdansk, Poland.

Moon I. 1994. *Modeling programmable logic controllers for logic verification*. IEEE Control Systems, 14**,** 2, pp. 53-59

Otter M., Årzén K., Dressler I., 2005 StateGraph - *A Modelica Library for Hierarchical State Machines*. Modelica 2005 Proceedings.

Roussel M., Denis B., 2002. *Safety properties verification of ladder diagram programs*. Journal Européen des Systèmes Automatisés, vol. 36, pp. 905-917

Seabra E., Machado J., 2007. Simulation of Real Time Systems Behavior Considering Human-Machine Interface. In Proceedings of the *6th EUROSIM Congress on Modelling and Simulation, Federation of European Simulation Societies*, September 9-13, Ljubljana, Slovenia.

Sheridan, T. B., 1984. Supervisory Control of Remote Manipulators Vehicules and Dynamic Processes: Experiments in Command and Display Aiding, In *Advances in Man-Machine Researches,* Vol.1.

# SLIDING MODE CONTROL
## *Is it Necessary Sliding Motion?*

L. Acho
*Department of Applied Mathematics III, EUETIB-Universidad Politécnica de Cataluña*
*Comte d'Urgell 187, 08036-Barcelona, Spain*
*leonardo.acho@upc.edu*

Keywords: Sliding Mode Control.

Abstract: Sliding mode control has been recognized to be insensitive to exogenous perturbations if the reachability conditions is warranted. Two phases follow the dynamics of the closed-loop perturbed system: 1) Finite-time convergence to the sliding surface, and 2) Sliding motion along the sliding surface. In the sliding motion the system has a reduced-order dynamic behavior. But, is it really necessary to have sliding motion to warranty the robustness property of the sliding mode controller? The main objective of this position paper is to theorize this important question.

## 1 INTRODUCTION

The most distinguished feature of sliding mode control it its ability to issue very robust control systems facing exogenous perturbations. Moreover, sliding mode control has been applied to a wide variety of control objectives such as regulation, tracking control, model following, adaptive control, observer design, among others (Perruqueti and Barbot, 2002; Edwards and Spurgeon, 1998). However, in all issues, sliding motion is secure to keep operable the sliding mode controller. Basically, the sliding controller involves two steps design. Ones is the finite-time convergence to the discontinuity manifold (the sliding surface), and the second one is the design of the sliding dynamics. However, and according with the examples given here, sliding motion is not necessary to warranty the main property of sliding controllers: insensitiveness to external perturbations. The outline of this position paper is as follows. The basic idea of sliding mode control using a second-order system is studied in Section two. Section three, two examples where not sliding motion exits all the time are granted and applied to a perturbed system illustrating that the insensitiveness property is not loss. Finally, Section five the conclusions are stated.

## 2 SLIDING MODE CONTROL

The basic idea of sliding mode control can be illustrated using second-order system (J. Y. Hung and HUng, 1993; Edwards and Spurgeon, 1998; Perruqueti and Barbot, 2002). At this point, consider the following system (Perruqueti and Barbot, 2002):

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= -x_2 + u. \end{aligned} \tag{1}$$

This system represents a dc-motor model, where $u$ is the scalar input control. Let us assume that the sliding surface is specified as:

$$s(x_1, x_2) = x_2 + \alpha x_1 = 0, \quad \alpha > 0. \tag{2}$$

The control objective consists to find a control law $u$ such that the sliding surface is attractive and reachable in finite time. In the sliding surface, the sliding mode motion then takes place (Perruqueti and Barbot, 2002; Edwards and Spurgeon, 1998; J. Y. Hung and HUng, 1993). The reachability in finite time is warranted if

$$s\dot{s} \leq \eta |s|, \quad \eta > 0, \tag{3}$$

called the $\eta-$reachability condition (Edwards and Spurgeon, 1998; Perruqueti and Barbot, 2002). Then, from (2), we have

$$s\dot{s} = s(\dot{x}_2 + \alpha \dot{x}_1). \tag{4}$$

Invoking (1), we get

$$\begin{aligned} s\dot{s} &= s(-x_2+u+\alpha x_2) \\ &= s((\alpha-1)x_2+u). \end{aligned} \quad (5)$$

If we set

$$u = -(\alpha-1)x_2 - \eta sgn(s), \quad \eta > 0, \quad (6)$$

we arrive to

$$s\dot{s} = s(-\eta sgn(s)) = -\eta|s|. \quad (7)$$

So, any trajectory of the closed-loop system (1) and (6), reaches the sliding surface $s = 0$ in finite time. Furthermore, in the sliding surface, the dynamic motion yields:

$$x_2 + \alpha x_1 = \dot{x}_1 + \alpha x_1 = 0 \Rightarrow \dot{x}_1 = -\alpha x_1 \quad (8)$$

which is asymptotically stable because $\alpha > 0$. Observe that the trajectory in the sliding surface can not scape from it. In resume, slide mode control involves two steps. Design a control law such as the sliding manifold be attractive and reachable in finite time, and design the sliding mode dynamics such as it is asymptotically stable. Also, it is recognizable that in sliding motion the order of the system is reduced from two to one. Let us now assume exogenous perturbation in our system. So, let it be as:

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= -x_2 + u + sin(t) \end{aligned} \quad (9)$$

where the exogenous perturbation is bounded. Then, using (6), we have:

$$s\dot{s} = -\eta|s| + sin(t)s \leq -\eta|s| + |s| = -|s|(\eta - 1). \quad (10)$$

So, the $\eta$-reachability condition is satisfied with $\eta > 1$ and any trajectory of the closed-loop perturbed system (9) and (6) reaches the sliding surface in finite time, where its sliding mode dynamics ($\dot{x}_1 = -\alpha x_1$) is unaltered in spite of the external perturbation. Simulations results are given in Fig. 1 with $\alpha = 0.5$ and $\eta = 2$.

# 3 MODIFIED SLIDING MODE CONTROL

Here, our main contributions are stated. The objective is to eliminate the persistency of chattering in the control law keeping the benefits of the sliding

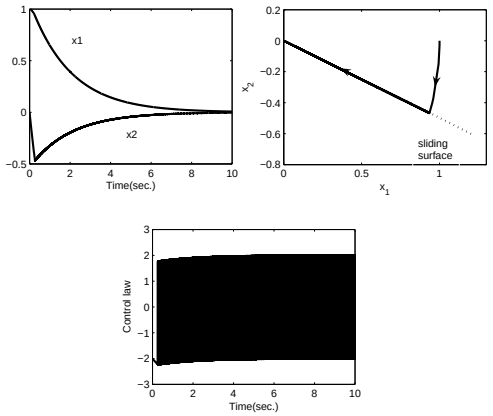



Figure 1: Simulations results of the perturbed system with $x_1(0) = 1$ and $x_2(0) = 0$.

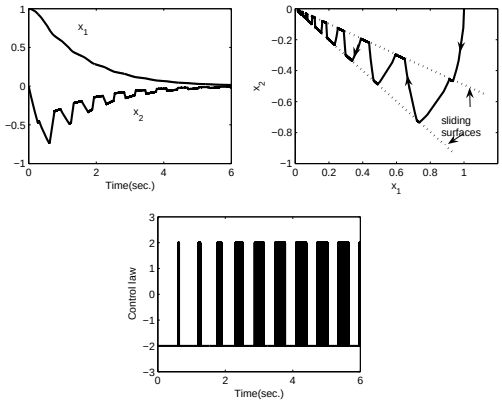



Figure 2: Simulations results of the perturbed system with $x_1(0) = 1$ and $x_2(0) = 0$ utilizing the modified sliding mode control.

mode control. From Fig. 1, when the system is perturbed and the $\eta$-reachability conditions is satisfied, the closed-loop system is insensitive to external perturbations, but chattering is sustained almost all the time. One way to reduce persistency on chattering is by means of changing the value of $\alpha$ in (6) and (2). For instance, employing the same perturbed system (9), with $\eta = 2$, and the same initial conditions; changing $\alpha$ from 0.5 to 1 at a frequency of $5/\pi Hz$ (a square signal), the simulation results are shown in Fig. 2. Here, the property of insensitiveness to external perturbations is evident and the chattering persistency is reduced. The commutation in the value of $\alpha$ is similar to commuting between two sliding surfaces (see Fig. 2). This works because the $\eta$-reachability condition ensures insensitiveness to exogenous perturbations. Here, sliding motion just occurs a certain time. So, the price is that sliding motion is not preserved

all the time; i.e., the reduced order dynamic motion is not sustained all the time. But, it is not so important from the robust control of view (mitigation of the external perturbation). Moreover, we can commutate between these sliding surfaces avoiding sliding motion; i.e., when the trajectory hits a sliding surface (warranted by the $\eta$-reachability condition), commutate to the other one (see Fig. 3). In Fig. 3, chattering occurs as a Zeno behavior, that is, the number of switching in the control law tends to infinite in finite time.

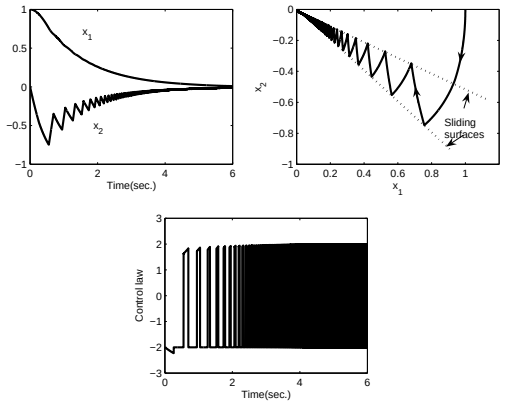



Figure 3: Simulations results of the perturbed system with $x_1(0) = 1$ and $x_2(0) = 0$ utilizing a second modified sliding mode control.

# 4 CONCLUSIONS

Employing second-order dynamic systems, we evidenced that sliding motion is not required to keeping insensitive property of sliding mode control. Also, with the examples shown here, chattering is not persistently exhibited. So, we have presented an important observation about this topic, that we think it will open further comments on it.

# REFERENCES

Edwards, C. and Spurgeon, S. K. (1998). *Sliding Mode Control: Theory and Applications*. Taylor and Francis, Ltd., UK, 1nd edition.

J. Y. Hung, W. G. and HUng, J. C. (1993). Variable structure control: A survey. In *IEEE Trans. on Industrial Electroncis*. Vol. 40, No.1, 2–2.

Perruqueti, W. and Barbot, J. P. (2002). *Sliding Mode Control in Engineering*. Marcel Dekker, Inc., New York, 1nd edition.

## AUTHOR INDEX

## AUTHOR INDEX (CONT.)

Proceedings of ICINCO 2008
Fifth International Conference on Informatics in Control, Automation and Robotics
ISBN: 978-989-8111-32-6
http://www.icinco.org