

Luis M.
Camarinha-Matos
(Ed.)



Technological Innovation for Sustainability

Second IFIP WG 5.5/SOCOLNET Doctoral Conference on
Computing, Electrical and Industrial Systems, DoCEIS 2011
Costa de Caparica, Portugal, February 2011
Proceedings

Editor-in-Chief

A. Joe Turner, Seneca, SC, USA

Editorial Board

Foundations of Computer Science

Mike Hinchey, Lero, Limerick, Ireland

Software: Theory and Practice

Bertrand Meyer, ETH Zurich, Switzerland

Education

Arthur Tatnall, Victoria University, Melbourne, Australia

Information Technology Applications

Ronald Waxman, EDA Standards Consulting, Beachwood, OH, USA

Communication Systems

Guy Leduc, Université de Liège, Belgium

System Modeling and Optimization

Jacques Henry, Université de Bordeaux, France

Information Systems

Jan Pries-Heje, Roskilde University, Denmark

Relationship between Computers and Society

Jackie Phahlamohlaka, CSIR, Pretoria, South Africa

Computer Systems Technology

Paolo Prinetto, Politecnico di Torino, Italy

Security and Privacy Protection in Information Processing Systems

Kai Rannenberg, Goethe University Frankfurt, Germany

Artificial Intelligence

Tharam Dillon, Curtin University, Bentley, Australia

Human-Computer Interaction

Annelise Mark Pejtersen, Center of Cognitive Systems Engineering, Denmark

Entertainment Computing

Ryohei Nakatsu, National University of Singapore

IFIP – The International Federation for Information Processing

IFIP was founded in 1960 under the auspices of UNESCO, following the First World Computer Congress held in Paris the previous year. An umbrella organization for societies working in information processing, IFIP's aim is two-fold: to support information processing within its member countries and to encourage technology transfer to developing nations. As its mission statement clearly states,

IFIP's mission is to be the leading, truly international, apolitical organization which encourages and assists in the development, exploitation and application of information technology for the benefit of all people.

IFIP is a non-profitmaking organization, run almost solely by 2500 volunteers. It operates through a number of technical committees, which organize events and publications. IFIP's events range from an international congress to local seminars, but the most important are:

- The IFIP World Computer Congress, held every second year;
- Open conferences;
- Working conferences.

The flagship event is the IFIP World Computer Congress, at which both invited and contributed papers are presented. Contributed papers are rigorously refereed and the rejection rate is high.

As with the Congress, participation in the open conferences is open to all and papers may be invited or submitted. Again, submitted papers are stringently refereed.

The working conferences are structured differently. They are usually run by a working group and attendance is small and by invitation only. Their purpose is to create an atmosphere conducive to innovation and development. Refereeing is less rigorous and papers are subjected to extensive group discussion.

Publications arising from IFIP events vary. The papers presented at the IFIP World Computer Congress and at open conferences are published as conference proceedings, while the results of the working conferences are often published as collections of selected and edited papers.

Any national society whose primary activity is in information may apply to become a full member of IFIP, although full membership is restricted to one society per country. Full members are entitled to vote at the annual General Assembly, National societies preferring a less committed involvement may apply for associate or corresponding membership. Associate members enjoy the same benefits as full members, but without voting rights. Corresponding members are not represented in IFIP bodies. Affiliated membership is open to non-national societies, and individual and honorary membership schemes are also offered.

Luis M. Camarinha-Matos (Ed.)

Technological Innovation for Sustainability

Second IFIP WG 5.5/SOCOLNET Doctoral Conference on
Computing, Electrical and Industrial Systems, DoCEIS 2011
Costa de Caparica, Portugal, February 21-23, 2011
Proceedings



Springer

Volume Editor

Luis M. Camarinha-Matos
New University of Lisbon
Faculty of Sciences and Technology
Campus de Caparica, 2829-516 Monte Caparica, Portugal
E-mail: cam@uninova.pt

ISSN 1868-4238

ISBN 978-3-642-19169-5

DOI 10.1007/978-3-642-19170-1

Springer Heidelberg Dordrecht London New York

e-ISSN 1868-422X

e-ISBN 978-3-642-19170-1

Library of Congress Control Number: 2011920513

CR Subject Classification (1998): C.2, H.1, C.4, I.2.9, C.3, J.2

© International Federation for Information Processing 2011

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

The use of general descriptive names, registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Technological Innovation and Sustainability Concerns

The subject of sustainability is a concern of growing importance, present in most strategic and political agendas, and also a prevalent issue in science and technology, leading to related terms such as sustainable development and even sustainability science. Encompassing a growing awareness of the political sectors and society in general for the importance of sustainability, the business sector has also started to acknowledge that preserving the environment and the other inter-related pillars of sustainability, i.e., the economic and social dimensions, is both good business and a moral obligation. New technological developments, in all fields, as major drivers of change, need to embed such concerns as well. As doctoral programs in science and engineering are important sources of innovative ideas and techniques that might lead to new products, technological innovation, and even new organizational and governance models with strong economic impact, it is important that the issue of sustainability becomes an intrinsic part of those programs.

Typically, PhD students are not experienced researchers, being rather in the process of learning how to do research. Nevertheless, a number of empiric studies also show that a high number of technological innovation ideas are produced in the early careers of researchers. From the combination of the eagerness to try new approaches and directions of young doctoral students with the experience and broad knowledge of their supervisors, an important pool of innovation potential emerges. The DoCEIS series of doctoral conferences on Computing, Electrical and Industrial Systems aim at creating a space for sharing and discussing ideas and results from doctoral research in these inter-related areas of engineering. Innovative ideas and hypotheses can be better enhanced when presented and discussed in an encouraging and open environment. DoCEIS aims to provide such an environment, releasing PhD students from the pressure of presenting their propositions in more formal contexts.

The second edition of DoCEIS, which was sponsored by SOCOLNET, IFIP and the IEEE Industrial Electronics Society, attracted a considerable number of paper submissions from a large number of PhD students (and their supervisors) from 16 countries. This book comprises the works selected by the International Program Committee for inclusion in the main program and covers a wide spectrum of topics, ranging from collaborative enterprise networks to microelectronics. Thus, novel results and ongoing research are presented, illustrated, and discussed in areas such as:

- Collaborative networks models and support
- Service-oriented systems
- Computational intelligence

- Robotic systems
- Petri nets
- Fault-tolerant systems
- Systems modelling and control
- Sensorial perception and signal processing
- Energy systems and novel electrical machinery

As a gluing element, all authors were asked to explicitly indicate the (potential) contribution of their work to sustainability.

We expect that this book will provide readers with an inspiring set of promising ideas, presented in a multi-disciplinary context, and that by their diversity these results can trigger and motivate new research and development directions.

We would like to thank all the authors for their contributions. We also appreciate the dedication of the DoCEIS Program Committee members who both helped with the selection of articles and contributed with valuable comments to improve their quality.

December 2010

Luis M. Camarinha-Matos



**Second IFIP / SOCOLNET Doctoral Conference on
COMPUTING, ELECTRICAL AND INDUSTRIAL
SYSTEMS**

Costa de Caparica, Portugal, February 21 – 23, 2011

Conference and Program Chair

Luis M. Camarinha-Matos (Portugal)

Program Committee

Hamideh Afsarmanesh (The Netherlands)	Stephan Kassel (Germany)
Jose Aguado (Spain)	Bernhard Katzy (Germany)
Amir Assadi (USA)	Marian Kazmierkowski (Poland)
José Barata (Portugal)	Tomasz Janowski (Macau)
Arnaldo Batista (Portugal)	Ricardo Jardim-Gonçalves (Portugal)
Luis Bernardo (Portugal)	Pontus Johnson (Sweden)
Xavier Boucher (France)	Paulo Leitão (Portugal)
Erik Bruun (Denmark)	J. Tenreiro Machado (Portugal)
Giuseppe Buja (Italy)	João Martins (Portugal)
António Cardoso (Portugal)	Maria do Carmo Medeiros (Portugal)
João Catalão (Portugal)	Paulo Miyagi (Brazil)
Wojciech Cellary (Poland)	Jörg Müller (Germany)
Jean-Jacques Chaillout (France)	Horacio Neto (Portugal)
David Chen (France)	Rui Neves-Silva (Portugal)
Fernando J. Coito (Portugal)	Mauro Onori (Sweden)
Ilhami Colak (Turkey)	Manuel D. Ortigueira (Portugal)
Luis Correia (Portugal)	Angel Ortiz (Spain)
Carlos Couto (Portugal)	Luis Palma (Portugal)
José Craveirinha (Portugal)	Willy Picard (Poland)
Jorge Dias (Portugal)	Paulo Pinto (Portugal)
H. Bulent Ertan (Turkey)	Ricardo Rabelo (Brazil)
Ip-Shing Fan (UK)	Hubert Razik (France)
Florin G Filip (Romania)	Sven-Volker Rehm (Germany)
Tarek Hassan (UK)	Rita Ribeiro (Portugal)
Maria Helena Fino (Portugal)	Juan Jose Rodriguez (Spain)
José M. Fonseca (Portugal)	Enrique Romero (Spain)
João Goes (Portugal)	José de la Rosa (Spain)
Luis Gomes (Portugal)	Luis Sá (Portugal)
Antoni Grau (Spain)	Gheorghe Scutaru (Romania)
Jose Luis Huertas (Spain)	Fernando Silva (Portugal)
Emmanouel Karapidakis (Greece)	Adolfo Steiger Garção (Portugal)
	Klaus-Dieter Thoben (Germany)

VIII Organization

Stanimir Valtchev (Portugal)
Manuela Vieira (Portugal)

Christian Vogel (Austria)
Antonio Volpentesta (Italy)

Organizing Committee Co-chairs

Luis Gomes (Portugal), João Goes (Portugal), João Martins (Portugal)

Organizing Committee (PhD Students)

António Pombo	José Lima
Eduardo Pinto	Carla Gomes
Fernando Pereira	Elena Baikova
Filipe Moutinho	Francisco Ganhão
José Lucas	Miguel Carneiro
Raul Cordeiro	Vitor Fialho
Svetlana Chemetova	Arnaldo Gouveia
Eduardo Eusébio	Dora Gonçalves
Joao Costa	Edinei Santin
José Luzio	João Casaleiro
José Silva	Manuel Vieira
Nuno Cardoso	

Technical Sponsors



Society of Collaborative Networks



IFIP WG 5.5 COVE
Co-Operation infrastructure for Virtual Enterprises
and electronic business



IEEE-Industrial Electronics Society

Organizational Sponsors



Organized by: PhD Program on Electrical and Computer Engineering FCT-UNL.

In collaboration with the PhD Program in Electrical and Computer Engineering - FCT-U Coimbra.

Table of Contents

Part 1: Collaborative Networks - I

Modelling Information Flow for Collaboration	3
<i>Christopher Durugbo, Ashutosh Tiwari, and Jeffrey R. Alcock</i>	
Value Systems Management Model for Co-innovation	11
<i>Patrícia Macedo and Luis M. Camarinha-Matos</i>	
ProActive Service Entity Framework: Improving Service Selection Chances within Large Senior Professional Virtual Community Scenario	21
<i>Tiago Cardoso and Luis M. Camarinha-Matos</i>	

Part 2: Collaborative Networks - II

Competence Mapping through Analysing Research Papers of a Scientific Community	33
<i>Antonio P. Volpentesta and Alberto M. Felicetti</i>	
Tuple-Based Semantic and Structural Mapping for a Sustainable Interoperability	45
<i>Carlos Agostinho, João Sarraipa, David Goncalves, and Ricardo Jardim-Goncalves</i>	
Planning and Scheduling for Dispersed and Collaborative Productive System	57
<i>Marcosiris A.O. Pessoa, Fabrício Junqueira, Paulo E. Miyagi, and Diolino J. Santos Fo</i>	

Part 3: Service-Oriented Systems

An SOA Based Approach to Improve Business Processes Flexibility in PLM	67
<i>Safa Hachani, Lilia Gzara, and Hervé Verjus</i>	
A Model for Automated Service Composition System in SOA Environment	75
<i>Paweł Stelmach, Adam Grzech, and Krzysztof Juszczyszyn</i>	
The Proposal of Service Oriented Data Mining System for Solving Real-Life Classification and Regression Problems	83
<i>Agnieszka Prusiewicz and Maciej Zięba</i>	

Personalisation in Service-Oriented Systems Using Markov Chain Model and Bayesian Inference	91
<i>Jakub M. Tomczak and Jerzy Świątek</i>	

Part 4: Computational Intelligence

Automatic Extraction of Document Topics	101
<i>Luís Teixeira, Gabriel Lopes, and Rita A. Ribeiro</i>	
Adaptive Imitation Scheme for Memetic Algorithms	109
<i>Ehsan Shahamatnia, Ramin Ayanzadeh, Rita A. Ribeiro, and Saeid Setayeshi</i>	
Design and Applications of Intelligent Systems in Identifying Future Occurrence of Tuberculosis Infection in Population at Risk	117
<i>Adel Ardalan, Ebru Selin Selen, Hesam Dashti, Adel Talaat, and Amir Assadi</i>	

Part 5: Robotic Systems - I

Gait Intention Analysis for Controlling Virtual Reality Walking Platforms	131
<i>Laura Madalina Dascalu, Adrian Stavar, and Doru Talaba</i>	
A Survey on Multi-robot Patrolling Algorithms	139
<i>David Portugal and Rui Rocha</i>	
Autonomous Planning Framework for Distributed Multiagent Robotic Systems	147
<i>Marko Švaco, Bojan Šekoranja, and Bojan Jerbić</i>	

Part 6: Robotic Systems - II

Controlling a Robotic Arm by Brainwaves and Eye Movement	157
<i>Cristian-Cezar Postelnicu, Doru Talaba, and Madalina-Ioana Toma</i>	
Robot Emotional State through Bayesian Visuo-Auditory Perception ...	165
<i>José Augusto Prado, Carlos Simplicio, and Jorge Dias</i>	
Manipulative Tasks Identification by Learning and Generalizing Hand Motions	173
<i>Diego R. Faria, Ricardo Martins, Jorge Lobo, and Jorge Dias</i>	
Evaluation of the Average Selection Speed Ratio between an Eye Tracking and a Head Tracking Interaction Interface	181
<i>Florin Bărbuceanu, Mihai Duguleană, Stoianovici Vlad, and Adrian Nedelcu</i>	

Part 7: Robotic Systems - III

LMA-Based Human Behaviour Analysis Using HMM	189
<i>Kamrad Khoshhal, Hadi Aliakbarpour, Kamel Mekhnacha, Julien Ros, João Quintas, and Jorge Dias</i>	
Daily Activity Model for Ambient Assisted Living	197
<i>GuoQing Yin and Dietmar Bruckner</i>	
Diagnosis in Networks of Mechatronic Agents: Validation of a Fault Propagation Model and Performance Assessment	205
<i>Luis Ribeiro, José Barata, Bruno Alves, and João Ferreira</i>	
Distributed Accelerometers for Gesture Recognition and Visualization	215
<i>Pedro Trindade and Jorge Lobo</i>	

Part 8: Petri Nets

Towards Statecharts to Input-Output Place Transition Nets Transformations	227
<i>Rui Pais, Luís Gomes, and João Paulo Barros</i>	
Petri Net Based Specification and Verification of Globally-Asynchronous-Locally-Synchronous System	237
<i>Filipe Moutinho, Luís Gomes, Paulo Barbosa, João Paulo Barros, Franklin Ramalho, Jorge Figueiredo, Anikó Costa, and André Monteiro</i>	
Automatic Generation of Run-Time Monitoring Capabilities to Petri Nets Based Controllers with Graphical User Interfaces	246
<i>Fernando Pereira, Luis Gomes, and Filipe Moutinho</i>	
SysVeritas: A Framework for Verifying IOPT Nets and Execution Semantics within Embedded Systems Design	256
<i>Paulo Barbosa, João Paulo Barros, Franklin Ramalho, Luís Gomes, Jorge Figueiredo, Filipe Moutinho, Anikó Costa, and André Aranha</i>	

Part 9: Sensorial and Perceptual Systems

Automatic Speech Recognition: An Improved Paradigm	269
<i>Tudor-Sabin Topoleanu and Gheorghe Leonte Mogan</i>	
HMM-Based Abnormal Behaviour Detection Using Heterogeneous Sensor Network	277
<i>Hadi Aliakbarpour, Kamrad Khoshhal, João Quintas, Kamel Mekhnacha, Julien Ros, Maria Andersson, and Jorge Dias</i>	

Displacement Measurements with ARPS in T-Beams Load Tests	286
<i>Graça Almeida, Fernando Melicio, Carlos Chastre, and José Fonseca</i>	

Part 10: Sensorial Systems and Decision

Wireless Monitoring and Remote Control of PV Systems Based on the ZigBee Protocol	297
<i>V. Katsioulis, E. Karapidakis, M. Hadjinicolaou, and A. Tsikalakis</i>	
A Linear Approach towards Modeling Human Behavior	305
<i>Rui Antunes, Fernando V. Coito, and Hermínio Duarte-Ramos</i>	
Nonlinear-Fuzzy Based Design Actuator Fault Diagnosis for the Satellite Attitude Control System	315
<i>Alireza Mirzaee and Ahmad Foruzantabar</i>	

Part 11: Signal Processing

Vergence Using GPU Cepstral Filtering	325
<i>Luis Almeida, Paulo Menezes, and Jorge Dias</i>	
Motion Patterns: Signal Interpretation towards the Laban Movement Analysis Semantics	333
<i>Luís Santos and Jorge Dias</i>	
ARMA Modelling of Sleep Spindles	341
<i>João Caldas da Costa, Manuel Duarte Ortigueira, and Arnaldo Batista</i>	
Pattern Recognition of the Household Water Consumption through Signal Analysis	349
<i>Giovana Almeida, José Vieira, José Marques, and Alberto Cardoso</i>	

Part 12: Fault-Tolerant Systems

Survey on Fault-Tolerant Diagnosis and Control Systems Applied to Multi-motor Electric Vehicles	359
<i>Alexandre Silveira, Rui Esteves Araújo, and Ricardo de Castro</i>	
Design of Active Holonic Fault-Tolerant Control Systems	367
<i>Robson M. da Silva, Paulo E. Miyagi, and Diolino J. Santos Filho</i>	
Design of Supervisory Control System for Ventricular Assist Device	375
<i>André Cavalheiro, Diolino Santos Fo., Aron Andrade, José Roberto Cardoso, Eduardo Bock, Jeison Fonseca, and Paulo Eigi Miyagi</i>	

Multi-agent Topologies over WSANs in the Context of Fault Tolerant Supervision	383
<i>Gonçalo Nunes, Alberto Cardoso, Amâncio Santos, and Paulo Gil</i>	

Part 13: Control Systems

Switched Unfalsified Multicontroller	393
<i>Fernando Costa, Fernando Coito, and Luís Palma</i>	
Design, Test and Experimental Validation of a VR Treadmill Walking Compensation Device	402
<i>Adrian Stavar, Laura Madalina Dascalu, and Doru Talaba</i>	
Design, Manufacturing and Tests of an Implantable Centrifugal Blood Pump	410
<i>Eduardo Bock, Pedro Antunes, Beatriz Uebelhart, Tarcísio Leão, Jeison Fonseca, André Cavalheiro, Diolino Santos Filho, José Roberto Cardoso, Bruno Utiyama, Juliana Leme, Cibele Silva, Aron Andrade, and Celso Arruda</i>	

Part 14: Energy Systems - I

Embedded Intelligent Structures for Energy Management in Vehicles ...	421
<i>Ana Puşcaş, Marius Carp, Paul Borza, and Iuliu Szekely</i>	
Energy Management System and Controlling Methods for a LDH1250HP Diesel Locomotive Based on Supercapacitors	429
<i>Marius Cătălin Carp, Ana Maria Puşcaş, and Paul Nicolae Borza</i>	
Home Electric Energy Monitoring System: Design and Prototyping	437
<i>João Gil Josué, João Murta Pina, and Mário Ventim Neves</i>	
Sustainable Housing Techno-Economic Feasibility Application	445
<i>Ricardo Francisco, Pedro Pereira, and João Martins</i>	

Part 15: Energy Systems - II

Study of Spread of Harmonics in an Electric Grid	457
<i>Sergio Ruiz Arranz, Enrique Romero-Cadaval, Eva González Romera, and María Isabel Milanés Montero</i>	
Impact of Grid Connected Photovoltaic System in the Power Quality of a Distribution Network	466
<i>Pedro González, Enrique Romero-Cadaval, Eva González, and Miguel A. Guerrero</i>	

Power Quality Disturbances Recognition Based on Grammatical Inference	474
---	-----

Tiago Fonseca and João F. Martins

Weather Monitoring System for Renewable Energy Power Production Correlation	481
---	-----

Marcos Afonso, Pedro Pereira, and João Martins

Part 16: Electrical Machines - I

Comparison of Different Modulation Strategies Applied to PMSM Drives under Inverter Fault Conditions	493
--	-----

Jorge O. Estima and A.J. Marques Cardoso

Optimization of Losses in Permanent Magnet Synchronous Motors for Electric Vehicle Application	502
--	-----

*Ana Isabel León-Sánchez, Enrique Romero-Cadaval,
María Isabel Milanés-Montero, and Javier Gallardo-Lozano*

A DC-DC Step-Up μ -Power Converter for Energy Harvesting Applications, Using Maximum Power Point Tracking, Based on Fractional Open Circuit Voltage	510
---	-----

Carlos Carvalho, Guilherme Lavareda, and Nuno Paulino

Wireless Sensor Network System for Measuring the Magnetic Noise of Inverter-Fed Three-Phase Induction Motors with Squirrel-Cage Rotor ...	518
---	-----

*Andrei Negoita, Gheorghe Scutaru, Ioan Peter, and
Razvan Mihai Ionescu*

Part 17: Electrical Machines - II

Axial Disc Motor Experimental Analysis Based in Steinmetz Parameters	529
--	-----

*David Inácio, João Martins, Mário Ventim Neves,
Alfredo Álvarez, and Amadeu Leão Rodrigues*

Transverse Flux Permanent Magnet Generator for Ocean Wave Energy Conversion	537
---	-----

José Lima, Anabela Pronto, and Mário Ventim Neves

A Fractional Power Disk Shaped Motor with Superconducting Armature	545
--	-----

*Gonçalo F. Luís, David Inácio, João Murta Pina, and
Mário Ventim Neves*

Numerical Design Methodology for an All Superconducting Linear Synchronous Motor	553
--	-----

João Murta Pina, Mário Ventim Neves, Alfredo Álvarez, and Amadeu Leão Rodrigues

Part 18: Electronics - I

CMOS Fully Differential Feedforward-Regulated Folded Cascode Amplifier	565
--	-----

Edinei Santin, Michael Figueiredo, João Goes, and Luís B. Oliveira

A New Modular Marx Derived Multilevel Converter	573
---	-----

Luis Encarnação, José Fernando Silva, Sónia F. Pinto, and Luis. M. Redondo

Energy Efficient NDMA Multi-packet Detection with Multiple Power Levels	581
---	-----

Francisco Ganhão, Miguel Pereira, Luis Bernardo, Rui Dinis, Rodolfo Oliveira, Paulo Pinto, Mário Macedo, and Paulo Pereira

Part 19: Electronics - II

Resistive Random Access Memories (RRAMs) Based on Metal Nanoparticles	591
---	-----

Asal Kiazadeh, Paulo R. Rocha, Qian Chen, and Henrique L. Gomes

Design, Synthesis, Characterization and Use of Random Conjugated Copolymers for Optoelectronic Applications	596
---	-----

Anna Calabrese, Andrea Pellegrino, Riccardo Po, Nicola Perin, Alessandra Tacca, Luca Longo, Nadia Camaioni, Francesca Tinti, Siraye E. Debebe, Salvatore Patanè, and Alfio Consoli

Optical Transducers Based on Amorphous Si/SiC Photodiodes	604
---	-----

Manuela Vieira, Paula Louro, Miguel Fernandes, Manuel A. Vieira, and João Costa

Author Index	613
---------------------------	-----

Modelling Information Flow for Collaboration

Christopher Durugbo, Ashutosh Tiwari, and Jeffrey R. Alcock

School of Applied Sciences, Cranfield University, MK43 0AL, United Kingdom
`{c.durugbo,a.tiwari,j.r.alcock}@cranfield.ac.uk`

Abstract. This paper describes a case study involving the use of a model of information flow, based on complex network theory, to study delivery information flow within a microsystem technology (MST) company. The structure of the company is captured through decision making, teamwork, and coordination networks. These networks were then aggregated and analysed from a sociological and an organisational perspective through measures that reflect connections, interconnectedness, and activity of individuals.

Keywords: information flow, complex networks, conceptual modelling, collaboration, delivery phase, microsystems technology.

1 Introduction

Recently, the study of social networks has offered a useful avenue for the analysis of information flow with a view to improving organisational factors and requirements such as capacity, productivity, efficiency, flexibility, and adaptability [1, 2]. Consequently, models of information flow of social networks have been proposed by analysts to explore minor and major roles of individuals in an organisation [3].

This research is motivated by the need to enhance collaboration during delivery phases for production through a model of information flow. Collaboration is an important organisational requirement that means working together in group(s) to achieve a common task or goal through teamwork, decision-making and coordination [4]. Enhancing collaboration during delivery offers opportunities for improving the efficiency, quality and sharing of information, leading to sustainable operations [4, 5].

The aim of this paper is to model complex networks for collaboration during delivery. In order to accomplish this, a case study of complex networks for the delivery of MST, will be undertaken to analyse information flow. §2 and §3 present the contribution of the paper to sustainability and the research background respectively, whereas §4 introduces a model of information flow used in §5 for a case study within an MST firm.

2 Contribution to Sustainability

This paper seeks to contribute to sustainable operations for organisations through a case study that makes use of a conceptual model of information flow to study the relationship between collaboration and delivery information flow. Sustainability is used here, from an economic perspective, to mean maintaining profitable operations.

3 Research Background

In [4], the state-of-the-art in the use of social network analysis (SNA) for modelling collaboration was analysed and the authors concluded that current models lacked visualisations for characterising formal relationships that symbolise collaboration roles and responsibilities. This is because SNA only considers individuals or groups of individuals as entities within social networks in terms of relationships, social roles and social structure [5, 6]. However, during collaborations, people/ teams are interconnected and tasks/processes are linked [4].

Interconnections for social networks have been the subject of research that have been used to model scale-free, hierarchical and random [7] properties and networks of organisations and communities. Social networks are also characterised by quantities that distinguish between structural interconnectedness and prominence of vertices. These quantities include [4]:

1. *Distance* - sum of links along the shortest path between vertices,
2. *Reachability* - establishes if vertices are linked directly or indirectly,
3. *Density* - compares number of actual links to possible links between vertices,
4. *Degree centrality* - number of vertices directly connected to a vertex,
5. *Closeness centrality* - inverse of the distance between a vertex and vertices in a network, and
6. *Betweenness centrality* - amount of times a vertex connects vertices to each other.

The need to include and analyse networks made up of tasks is evident in current studies by authors such as Batallas and Yassine [8], in which the analysis of social networks is complemented with design structure matrices for analysing tasks and Collins *et al.* [9] that examined task networks for product development. These studies have mainly concentrated on isolating and analysing social and task networks separately or making use of one technique to analyse the other. An inspection of these techniques suggests that potentially some links and flows may be omitted. For instance, a human operator working as part of a team may access or transfer some information necessary for collaboration (such as number of products to be manufactured) with a manufacturing process. This interaction may not require the participation of a team member or may be accessed by another team member through the process without a direct link to the original source of the information i.e. the first human participant.

4 A Model of Information Flow for Organisational Collaboration

The focus of this paper is to make use of a model of information flow, shown in Fig. 1, to analyse collaborative delivery in an MST company.

As described in [4], mathematically, a collaborating organisation, can be modelled as a connected, partitioned, non-overlapping hypergraph $G = (V, E)$ containing a graph for characterising the collaborative social network of individuals/groups $G_s = (V_s, E_s)$ and a directed graph for characterising the collaborative activity network of processes/tasks $G_p = (V_p, E_p)$. V_s represents social vertices of collaborating individuals, teams or organisations, and V_p represents activity vertices for processes that are required to achieve a common goal that could not be achieved by the independently

collaborating individuals. E_s and E_p correspond to edges between teams (or individuals) and processes. In the proposed model, processes become part of a collaboration based on the set of interface edges T showing connections between human participants and processes i.e. T associates V_s with V_p . Consequently G is defined by $V = V_s \cup V_p$ and $V_s \cap V_p = \emptyset$. Similarly, $E = E_s \cup E_p \cup T$ and $E_s \cap E_p \cap T = \emptyset$.

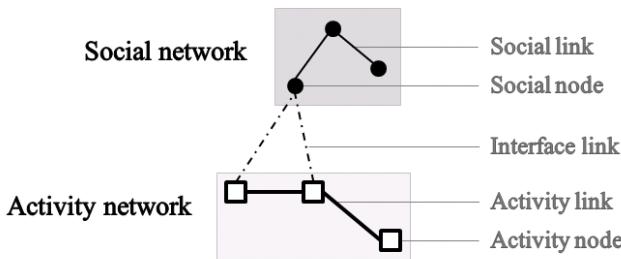


Fig. 1. A model of information flow for organisational collaboration [4]

Using the proposed information structure, the information behaviour for organisations can be characterised using key SNA measures of clustering coefficient, closeness and degree centrality. These quantities were selected because they reflect interconnectedness within groups, individual connections for relationships and activity of individuals respectively [5, 6].

The **degree centrality** (Dc_i) is a ratio of number of directly connected vertices to the number of possible vertices in a network and can be computed as:

$$Dc_i = [\deg]_i / N - 1 \quad (1)$$

Where, N is the number of vertices in the network and $[\deg]_i$ is the number of vertices directly connected to i .

The **clustering coefficient** assesses the density between vertices and represents the tendency for vertices to cluster together. If a vertex i , connects to b_i neighbours, and the number of possible edges between the vertices is given as $b_i(b_i - 1)/2$, then the clustering coefficient (Cc_i) of i can be computed as:

$$Cc_i = 2n_i / b_i(b_i - 1) \quad (2)$$

Where n_i is the number of edges between b_i neighbours.

The **closeness** between vertices defines the order with which one vertex collaborates with another vertex. It is computed as the inverse of the geodesic distance (d_{ij}) between a pair of vertices i and j . d_{ij} is the number of edges along the shortest path between i and j . Closeness (c_{ij}) can be calculated as:

$$c_{ij} = 1 / \sum_{i \neq j \in N} d_{ij} \quad (3)$$

For instance, if a vertex i connects directly to another vertex j , then the closeness of i to j is given as 1, if collaboration is established as a result of connecting to a third vertex k acting as a hub i.e. dictator collaboration [4], then i has a closeness of 0.5 to j .

5 Case Study

Company B is an MST company based in the United Kingdom with a targeted global market. It operates with 14 staff for the delivery of microfluidic and microoptical based products and services as business-to-business solutions for customers that are mainly original equipment manufacturers or an academic institution. Products delivered by Company B include microlens arrays for flat panel displays, and lab-on-a-chip microfluidic devices for industrial automation, cell analysis and drug delivery.

In this section, the research method and findings of the case study to analyse delivery information flow in Company B is described. The implications of the findings for sustainable operations are also identified.

5.1 Research Methodology

An analytical research methodology [10] was adopted for the case study in two steps.

Firstly, the information structure of Company B was analysed through an initial semi-structured telephone interview with the customer support manager (CSM) at Company B that lasted 25 minutes. This interviewee was provided by the company director at Company B following initial telephone conversations to request permission to carry out the study. The director designated the CSM as personnel responsible for managing the flow of information during MST delivery and the main question posed to the CSM to initiate the semi-structured interviews was: ‘What are the processes and information flow for the delivery phase in your company?’

Secondly, using the data provided by the CSM, face-to-face interviews were then conducted on-site with 7 other available company personnel involved in the delivery process to analyse information flow using the proposed model. In order to populate the model, interviewees were asked to validate the description provided by the CSM. Interviewees were also asked to identify processes and other personnel that they were connected to during delivery for decision making, teamwork, and coordination.

5.2 Research Findings

The interviews with personnel at Company B revealed that decision-making within G_s is a directed graph that involves the entire social vertices within V_s . Incident edges are mainly outwards from the subset of V_s containing social vertices of the management group (Ba to Bd) as shown in Fig. 2a.

Teamwork for the social network for Company B is an undirected graph with clusters formed among groups of social vertices for V_s whereas coordination within Company B is also an undirected graph involving the entire network.

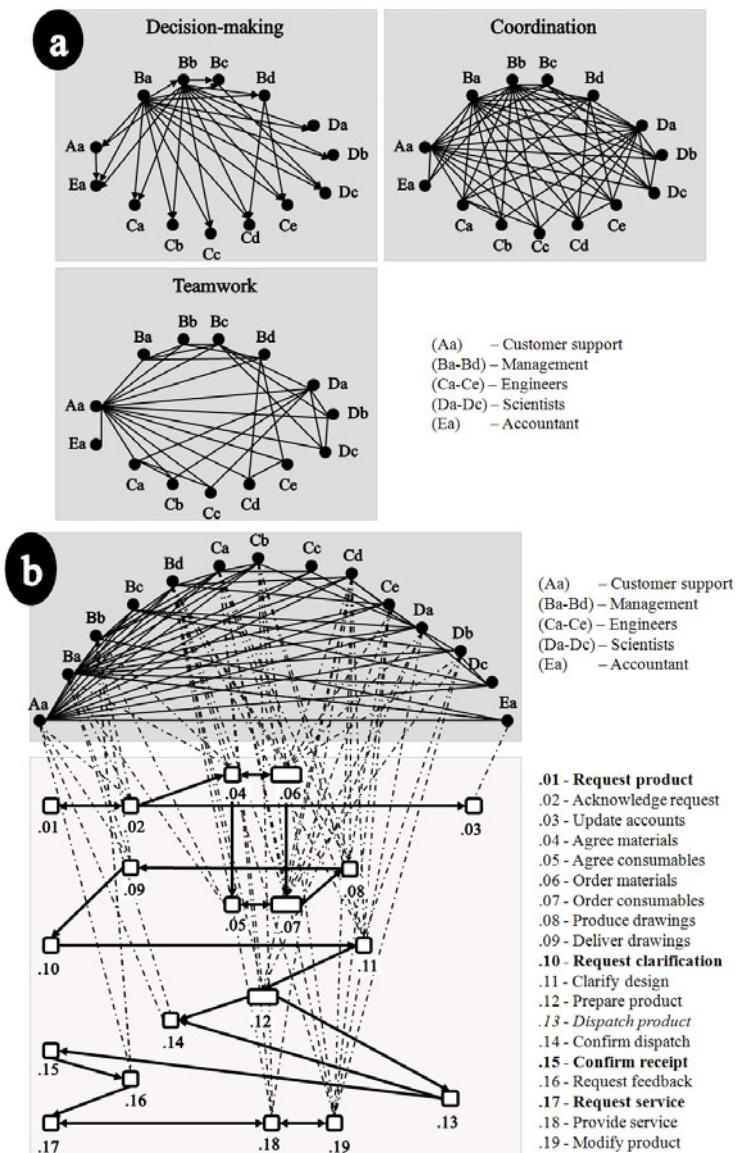


Fig. 2. (a) Social networks for decision-making, coordination and teamwork in Company B, and (b) Information structure for Company B

An aggregation of the social networks in Fig. 2a and the 19 (.01-19) processes described by the CSM (that was validated by the participants of the face-to-face interviews) resulted in the information structure for Company B, as shown in Fig. 2b. Activity vertices .01 (request product), .10 (request clarification), .15 (confirm receipt) and .17 (request service) are carried out by the customer whereas .13

(dispatch product) is the responsibility of the courier provider. For collaboration within Company A, the 14 available staff i.e. the social vertices Aa-Ea, are connected to the remaining 14 processes through 57 internal interface edges.

Management processes, associated with vertices Aa and Ba-Bd, form a subset of V_p consisting of activity vertices .02 (acknowledge request), .09 (deliver drawings), .14 (confirm dispatch) and .16 (request feedback). Similarly, engineering and science processes, associated with social vertices Ca-Ce and Da-Dc, form a subset of V_p involving activity vertices .04 (agree materials), .05 (agree consumables), .06 (order materials), .07 (order consumables), .08 (produce drawings), .11 (clarify design), .12 (prepare product), .18 (provide service) and .19 (modify product). The accounting process associated with Ea is activity vertex .03 (update account).

While the maximum number of edges in a fully connected social network (G_s) of 14 social vertices can be computed as $G_s(G_s - 1)/2$ i.e. 91, the maximum number of interface edges to the 14 activity vertices within G_p (.02-.09, .11, .12, .14, .16, .18 and .19) on which collaboration is based, can be calculated as $G_s \times G_p$ i.e. 196.

As shown in Fig 2b, vertex Aa connects to 13 social vertices and 4 activity vertices. Consequently the values of Dc_i for vertex Aa within the G_s and G can be computed from eqn. (1) as $13/(14 - 1) = 1.000$ and $(13+4)/(14+(14 - 1)) = 0.630$. Since, the social vertices directly connected to Aa form 54 edges with each other and 57 interface edges with directly connected activity vertices, the value of Cc_i for Aa within the social G_s and G can be calculated from eqn. (2) as $54/91 = 0.593$ and $(54+57)/(91+196) = 0.387$ respectively. Similarly, the values of c_{ij} within G_s and G can be determined from eqn. (3) as $1/(13 \times 1) = 0.077$ and $1/((13 \times 1) + ((4 \times 1) + ((19 - 4) \times 2))) = 0.021$. Values of Dc_i , Cc_i and c_{ij} for the social vertices in Company B have been computed using a similar approach and are shown in Table 1.

With an overall average Dc_i value of 0.593 (59.3%) and 0.437 (43.7%) for G_s and G from Table 1, this study suggests a significant level of interconnectedness within Company B. Similarly, the values of c_{ij} (0.057 and 0.019 for G_s and G) suggest

Table 1. Social network measures for social vertices within the social network (G_s) and entire network (G) of Company B (degree centrality - Dc_i , clustering coefficient - Cc_i , closeness - c_{ij})

Social vertices	Within G_s			Within G		
	Dc_i	Cc_i	c_{ij}	Dc_i	Cc_i	c_{ij}
Aa	1.000	0.593	0.077	0.630	0.387	0.021
Ba	1.000	0.593	0.077	0.667	0.387	0.022
Bb	1.000	0.593	0.077	0.667	0.387	0.022
Bc	0.462	0.231	0.050	0.222	0.150	0.017
Bd	0.538	0.286	0.053	0.593	0.220	0.020
Ca	0.538	0.275	0.053	0.519	0.226	0.020
Cb	0.385	0.165	0.048	0.407	0.160	0.019
Cc	0.385	0.165	0.048	0.185	0.139	0.017
Cd	0.462	0.231	0.050	0.481	0.185	0.019
Ce	0.462	0.231	0.050	0.407	0.209	0.019
Da	0.923	0.582	0.071	0.593	0.380	0.020
Db	0.462	0.231	0.050	0.370	0.150	0.017
Dc	0.462	0.231	0.050	0.222	0.150	0.017
Ea	0.231	0.033	0.043	0.148	0.073	0.015
Overall average	0.593	0.317	0.057	0.437	0.229	0.019

high activity of personnel within Company B. This is because if all Company B's personnel were fully active, i.e. each vertex can connect directly to other vertices, then the average values of c_{ij} for Company B would be 0.080 and 0.037 for G_s and G respectively. In Company B's current state, this represents 74.1% and 51.3% of potential activity within G_s and G respectively. However, low values of Cc_i (0.317 and 0.229 for G_s and G) indicate possible weak individual connections for relationships across the organisation. Nonetheless, the average value of Cc_i is higher when analysed for social vertices within working groups (such as activity vertices Da-Dc), indicating stronger connections for teams as opposed to the entire organisation. This finding correlates with existing studies in which it is suggested that small and medium enterprises (SMEs) within high-tech firms, such as Company B, are effective at working together for innovation [11].

5.3 Implications for Sustainable Operations

Two important lessons for sustainable operations were learnt from the case study involving considerations for dichotomies based on small-scale R&D (research and development) vs. large-scale manufacturing, and hierarchical vs. flat structures.

Firstly, although Company B applies a small-volume-large-variety production business model, on-going efforts to expand operations for large scale production poses major challenges for coordination, decision-making and teamwork. Further analysis following discussions with the company director of Company B revealed that this challenge is a major cause for disagreements and conflicts that has currently severed ties and friendships within the company, impacting on collaboration levels.

Secondly, further discussions with personnel also revealed a split between management staff wanting fewer organisational layers to ease the flow of information and engineering personnel favouring hierarchies for structure in processes. However, non-management personnel (such as vertex Da) have been able establish links across the divide through negotiation and interpersonal contact. In contrast, some experienced management personnel (such as Bc and Bd) have not been able to effectively collaborate in Company B, a situation which according to the company director is due to the split between R&D and manufacturing, and poor interpersonal skills. These lessons reinforce the importance of trade-offs for MST firms [12] that offer avenues for maintaining collaboration as well as business driven and simplified information flow.

6 Conclusions

In this paper, a model of information flow for collaboration is proposed as a combination of *social networks* involving decision-making, coordination and teamwork of human entities and *activity networks* containing non-human entities such as production processes, technologies and systems. The model also applies measures of degree centrality, clustering coefficient and closeness from social network analysis (SNA) to assess the level of collaboration within organisations. Useful insights from a case study of delivery within a microsystem technology (MST) firm, based on the proposed model, suggested that high-tech small and medium enterprises can be effective at collaborating for delivery. For the case study, social vertices (i.e. participants in collaborations) were assessed using the SNA measures from: a sociological

perspective for the social network and an organisational perspective for the entire (social and activity) network. Findings from the study also suggested that effective collaborative delivery requires trade-offs for strategising capabilities for operations and for managing organisational layers that enable expansion.

Acknowledgments. The authors would like to extend their sincere thanks to the Engineering and Physical Sciences Research Council (EPSRC), for its support via the Cranfield Innovative Manufacturing Research Centre (CIMRC), towards the work carried out in the preparation of this paper.

References

1. Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., Hwang, D.-U.: Complex networks: Structure and dynamics. *Phys. Rep.* 424(4-5), 175–308 (2006)
2. Schultz-Jones, B.: Examining information behavior through social networks: An interdisciplinary review. *J. Doc.* 65(4), 592–631 (2009)
3. White, L.: Connecting organizations: Developing the idea of network learning in inter-organizational settings. *Syst. Res. Behav. Sci.* 25(6), 701–716 (2008)
4. Durugbo, C., Hutabarat, W., Tiwari, A., Alcock, J.R.: Modelling Collaboration using Complex Networks. Submitted to *Inform. Sciences* (2011)
5. Durugbo, C., Tiwari, A., Alcock, J.R.: An Infodynamic Engine Approach to Improving the Efficiency of Information Flow in a Product-Service System. In: CIRP IPS2 Conf., pp. 107–112 (2009)
6. Hatala, J.-P., Lutta, J.G.: Managing information sharing within an organizational setting: A social network perspective. *Perform. Imp. Quart.* 21(4), 5–33 (2009)
7. Valente, T.W., Coronges, K.A., Stevens, G.D., Cousineau, M.R.: Collaboration and competition in a children's health initiative coalition: A network analysis. *Evalu. Program Plann.* 31(4), 392–402 (2008)
8. Barabasi, A.L., Oltvai, Z.N.: Network Biology: Understanding the Cell's Functional Organization. *Nat. Rev. Genet.* 5(2), 101–113 (2004)
9. Batallas, D.A., Yassine, A.A.: Information leaders in product development organizational networks: Social network analysis of the design structure matrix. *IEEE T. Eng. Manage.* 53(4), 570–582 (2006)
10. Collins, S.T., Bradley, J.A., Yassine, A.A.: Analyzing Product Development Task Networks to Examine Organizational Change. *IEEE T. Eng. Manage.* 57(3), 513–525 (2010)
11. Kumar, R.: Research methodology. Longman, London (1996)
12. Trumbach, C.C., Payne, D., Kongthon, A.: Technology mining for small firms: Knowledge prospecting for competitive advantage. *Technol. Forecast. Soc.* 73(8), 937–949 (2006)
13. Durugbo, C., Tiwari, A., Alcock, J.R.: Survey of media forms and information flow models in microsystems companies. In: Camarinha-Matos, L.M., Pereira, P., Ribeiro, L. (eds.) DoCEIS 2010. IFIP AICT, vol. 314, pp. 62–69. Springer, Heidelberg (2010)

Value Systems Management Model for Co-innovation

Patrícia Macedo^{1,2} and Luis M. Camarinha-Matos¹

¹ Faculty of Sciences and Technology, Universidade Nova de Lisboa, Portugal

² Polytechnic Institute of Setubal, Portugal

pmacedo@est.ips.pt, cam@uninova.pt

Abstract. Nowadays innovation and collaboration are strategic issues for enterprises to remain competitive in the global market. Many new developments are carried out with external partners, including universities and research institutions aiming to generate novel solutions in order to improve business performance and sustainability. However, the balance between intellectual property protection and intellectual property sharing is hard to manage. In order to increase the sustainability of innovation networks it is important to provide mechanisms to easily define the profile of the collaboration and to assess the degree of alignment with the potential partners. This paper aims to discuss how these mechanisms can be provided through the implementation of a Core Value Management System in the scope of co-innovation.

Keywords: collaborative networks, value systems, intellectual property, co-innovation.

1 Introduction

Nowadays, the strive to achieve innovation correspond to a very costly and risky process, being difficult for companies to innovate in short periods of time in a global market, where customers' needs change quickly and the products/services' life cycles get shorter. This is a particular tough challenge for small and medium enterprises (SMEs). Therefore, many new developments are carried out with external partners, including universities and research institutions with the aim to improve business performance and its sustainability, as well as to reduce risks.

Co-innovation is a new business paradigm where it is assumed that firms or individual persons can establish a partnership with the aim of jointly developing new ideas and new products [1]. These kinds of partnerships raise new challenges in the scope of trust management and intellectual property management [2, 3], since the balance between intellectual property sharing and intellectual property protection is hard to manage. Moreover, if partners have different perceptions of outcomes and different notions of the expected behavior, this might, in some cases, lead to some behaviors that compromise collaboration sustainability. For instance, if a firm does not believe in the importance of sharing knowledge, then it can be expected that its behavior will not contribute positively to the development and implementation of new ideas or new technologies in alliance with other firms.

Since a Value System defines the set of values and priorities that regulate the behavior of an organization, it determines or at least constrains the decision-making

processes of that organization. Therefore, the identification and characterization of the *Value Systems* of the network and its members is fundamental when attempting to improve collaboration [4]. Consequently it is important to develop mechanisms to facilitate the definition of the profile of the co-innovation partnership in terms of values. Moreover, it is also relevant to develop mechanism to assess the degree of alignment between the values required for the partnership and the values hold by the potential partners.

The research performed aimed to address the following question: *What would be an adequate modeling framework for supporting effectively the specification and analysis of Value System in collaborative environments?*

Even so, this paper presents only part of the results achieved during this research, more specifically, it discusses how the models and methods proposed in [5-7] can be applied to analyze and assess the level of alignment between the network and the potential partners in terms of collaborative innovation profile.

2 Contributions to Sustainability

The development of new products and services in a short period of time is a key condition to survive in the global market, however this represents also a big challenge for small and medium enterprises. The co-innovation paradigm brings a new way to conduct innovation, being focused on the cooperation with others to achieve innovations. In order to support this paradigm it is important to provide mechanisms to effortlessly define the values profile required for the innovation network and to assess the degree of values alignment between it and its potential partners. This paper aims at contributing to sustainability under the economic perspective in the way that it proposes mechanisms to deal with some of the challenges of co-innovation, thus leading to better survival of SMEs in turbulent market conditions.

3 Related Work on Value Systems Management

Early Value Systems management studies have been conducted essentially in social sciences, where scientists have discussed the importance of values management for the success of organizations and the importance of values alignment. For instance, Rokeach and Schwartz [8, 9] developed some empirical methods to identify core-values. Based on this work they proposed organizational core-values taxonomies. From the economic field perspective, the Value System concept has been developed based on the assumption that value means how much (usually money) a product or service is worth to someone. This notion has been introduced by Porter [10], that considered that Value System management comprises the management of the set of activity links where there is value creation, such as the links among a firm and its suppliers, other businesses within the firm's corporate family, distribution channels and the firm's end-user customers. This notion has been extended by Alle [11] towards supporting a wider notion of value under which the term is associated to "anything that can give rise to exchange", leading to the *Value Network Model*. In recent years some studies have explored the importance of Value Systems in the context of Collaborative Networks (CN) [4, 12, 13]. Furthermore, a set of tools (frameworks, methods and a

web application) [6, 7] have been proposed based on a Value System Conceptual Model to support the configuration and analysis of Value Systems in collaborative environments. However, such set of tools have just been applied to CNs in general terms, none of them being applied to support co-innovation management in particular.

4 Core Value System Management Model

In previous works, we have proposed a conceptual model for Value Systems [5] in an attempt to provide a unified definition of the concept that embraces not only the notion accepted in Sociology, but also the notions developed in the Economy and Knowledge Management fields for Value Systems.

A **Value** represents the relative importance of an entity, for a given evaluator.

A **Core-value** is a main characteristic of the organization (or CN) that motivates and regulates its behavior.

A **Value System** is defined by the set of things (resources, processes, behaviours, relationships, beliefs, information) that can be evaluated and that are important to the evaluator, as well as, the respective mechanisms of their evaluation.

A **Core Value System** is a specialization of Value System, that assumes the organization or the CN as the object of evaluation, and the core-values as the set of characteristics to be evaluated.

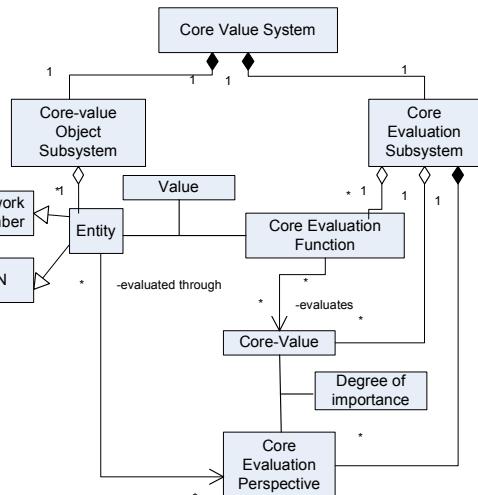


Fig. 1. UML Diagram of the CVS Conceptual Model

The set of characteristics that each organization (or network of organisations) considers as the most important for itself and that motivate or regulate its behavior are called *core-values*. Departing from the notion of core-values we have introduced a conceptual model for Core Value System (CVS) [6], that encompasses the core-values notion. This concept is a restricted view of the generic Value System model of which it can be considered a specialisation. The UML class diagram presented in Fig. 1 illustrates the structure of the CVS model. Accordingly, a CVS is composed of two subsystems: (i) the core-value objects subsystem, which is represented by the organization or networked organization itself; (ii) the core evaluation subsystem, which represents the elements of evaluation, such as the core-values, the core-evaluation perspective and the functions to evaluate the organization's core-values. The set of core-values of an actor and respective preferences (degree of importance) are represented according to the CVS conceptual model by the core-evaluation perspective, which becomes the main structural element in the proposed approach.

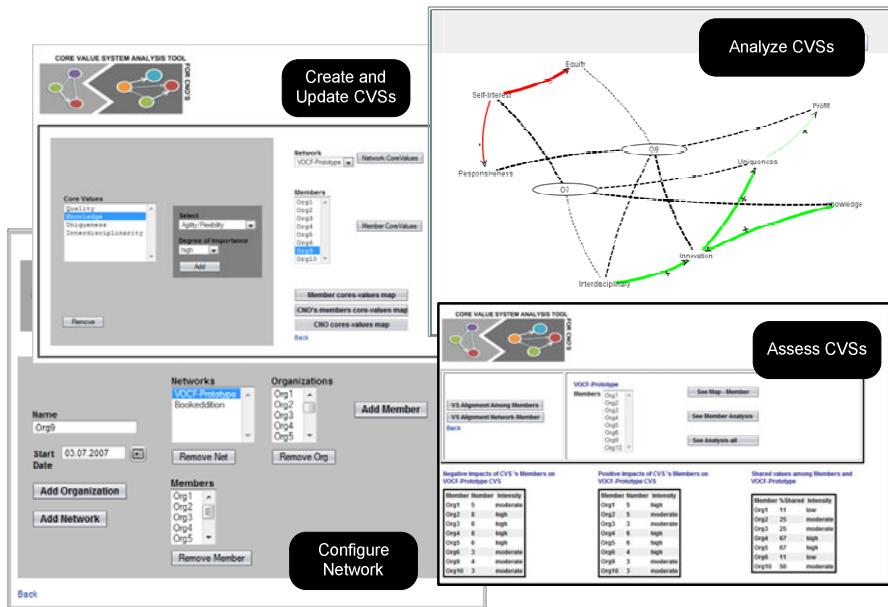


Fig. 2. Web-based tool to support CVSs Management

Aiming to provide methods to systematically analyze CVSs in collaborative environments, we have also proposed a set of qualitative reasoning methods [7] supported in a framework of analysis based on qualitative causal maps and graphs. The construction of three elementary maps is proposed in this framework.

1. *Core-values influence map:* Use of causal maps to show how core-values influence positively or negatively each other, and the intensity of the influence.
2. *Organizations' core-values map:* Use of graphs to show the core-values held by each organization and its degree of importance, as well as the core-values shared by organizations.
3. *CN's core-values map:* Use of graphs to show the core-values held by the CN and their degree of importance.

Departing from these elementary maps, it is possible to aggregate them in order to build maps that evidence the impact of one CVS onto another [6], facilitating the analysis process. Analysis is one of the components of the CVS management process, which includes the following activities:

- **Create:** CVS are configured. For each CVS, the core-evaluation perspectives have to be defined for the network and for each member.
- **Update:** CVS can be modified. As priorities can change during long-term collaboration activities, there is the need to adjust the degree of importance of the core-values.
- **Analyse:** The created CVS has to be examined in detail, in order to understand which core-values are shared with other partners, and which core-values influence positively or negatively other CVSs. This analysis should cover the network

level and the member level. At the network level is analyzed the network CVSs and the impact of the members' CVSs on the network's CVS are analyzed, while at the member level the CVS of each member and the interaction among members' CVSs are analyze.

- **Assess:** A comparative assessment can be made, where the level of fitness between two CVSs is assessed using distinct criteria. This assessment can also cover the network level and the member level, considering in each case the degree of importance of the core-values to infer the level of CVSs alignment.

A web-based tool to support the CVS management was developed (see Fig. 2). Essentially, the CVSs alignment analysis can be performed among network members' CVSs, or between the network's CVS and the CVS of a partner or a potential partner. The tool is divided into four functional components: (i) Core-values knowledge management – to be used by the knowledge experts in order to specify core-values and their characteristics; (ii) Core Value System Configuration – to be used by brokers, network managers or network members in order to define their CVSs; (iii) Core Value Systems Analysis – to be used by brokers, network managers and network members in order to analyze their CVS; (iv) Access management tool – provides features that allow the tool manager to configure accesses to the tool according to the user profiles.

5 Applying Value System Management to Support Co-Innovation

Theoretical considerations: The alliances established with the aim of developing new products and new technologies should have specific characteristic. Chesbrough [3] defends that for firms to be able to successfully implement the co-innovation business model, they have to be capable of sharing their knowledge, being flexible and being responsive. However, Flores et al. [2] defend another set of characteristics as being the most important ones for working in an co-innovation business model. Although there is no unique set of common characteristics accepted by the researchers, it is acknowledged that the set of core-values taken by each firm, determine their behaviour in the collaboration process of creation. Therefore, the definition of the set of characteristics required to work in a sustainable way in an innovation network, can be an important step for the sustainability of these networks. As such, the *collaborative innovation evaluation perspective* will represent this set of required characteristics. The *collaborative innovation evaluation* is in fact, an instantiation of a *core-evaluation perspective*, and is defined as: $pe_{co-innovation} = \{< dv, wv >\}$, where:

- dv is the vector of core-values considered as relevant for the co-innovation process. $dv = [cv_1, cv_2, \dots, cv_n]$: $cv_i \in CV$
- wv represents the weights-vector, where each element defines the degree of importance of the corresponding core-value. These preferences can be expressed qualitatively.

As previously proposed in [6] and [7], not just the common core-values will be considered as relevant criteria to assess the values alignment between two entities, but also the core-values that influence positively or negatively the core-values specified in the *collaborative innovation evaluation perspective* will be considered. Thus, the notion

of *collaborative innovation value profile* is introduced to cover this idea. The *collaboration innovation value profile* of a partner shows which of its core-values are aligned and which are misaligned with the collaborative innovation characteristics required for the partnership. That is indeed, all the core-values belonging to the partner's CVS that meet one of the following criteria:

- the core-value belongs to the collaborative innovation evaluation perspective.
- the core-value has a positive influence on a core-value belonging to the collaborative innovation evaluation perspective.
- the core-value has a negative influence on a core-value belonging to the collaborative innovation evaluation perspective.

The use of the mentioned set of maps based on causal maps and graph theory in the context of the innovation network, will allow us to easily identify the core-values that compose the *collaboration innovation value profile* of each network member. The three alignment indicators proposed to assess values alignment [7]: (i) Shared Values Level; (ii) Positive Impact Level; (iii) Potential for conflict level, may be applied to calculate the *Collaborative Innovation Values Profile* using just minor adjustments, such as:

- A *collaborative innovation evaluation perspective* has to be configured in the CVS of the innovation network.
- A *core-evaluation perspective* has to be configured in the CVS of each potential partner.
- Through the observation of the complete aggregate maps generated for the innovation network, the core-values belonging to each *Collaborative Innovation Values Profile* are identified.
- The *Shared Values Level*, the *Positive Impact Level*, and the *Potential for conflict level* between the network and each partner are calculated, obtaining the level of each core-value in the profile.

The example below illustrates how these concepts can be applied.

Illustrative example: This example intends to illustrate how to apply the Value System Management Model in order to assess the Collaborative Innovation Values Profile Alignment. The data used in this example was obtained through a survey conducted in the scope of a case study done inside the ECOLEAD project [14]. Hence, this example cannot be considered as a full case study, but as a potential application where the scenario and the data used are both real.

The ECOLEAD project was divided into several work packages. Each work package team was created to respond to a specific set of objectives. For this illustrative example the development of some components of the VO Creation Framework (VOCF) prototype is considered. The development of this task can be considered as a co-innovation process, since it comprises the collaborative development of an innovative product carried out by a heterogeneous group of organizations. For this group the following collaborative innovation evaluation *perspective* was specified:

$$\begin{aligned}
 p_{co-innovation} &<dv_I, wv_I> \\
 dv_I &= [Innovation, Knowledge Sharing, Agility, Responsiveness] \\
 wv_I &= [high, very high, high, very high].
 \end{aligned}$$

Two Universities, one Research Institute of Applied research, and one SME Network joined the group (for a question of privacy, the partners are not identified). For these four organizations the work package manager has provided the data required for specifying the following *core-evaluation perspectives*:

$pe_{univA} <= [innovation, knowledge, reputation, interdisciplinary, quality, knowledgeSharing], [very_high, very_high, fair, fair, high, high]$

$pe_{univB} <= [innovation, knowledge, reputation, interdisciplinary, reliability, knowledgeSharing], [very_high, very_high, fair, high, high, high]$

$pe_{researchInstitute} <= [Innovation, Uniqueness, Self-Interest, Interdisciplinary], [high, very_high, fair, high]$

$pe_{SMEnetwork} <= [Profit, FinancialStability, Responsiveness, Agility], [very_high, very_high, very_high, fair]$

Table 1. Core-values Alignment Assessment Results

	Shared Values Level	Positive Impact Level	Negative Impact Level
UnivA	Moderate	High	None
UnivB	Moderate	Moderate	None
Research Institute(RI)	Low	Low	High
SME Network	Low	Moderate	None

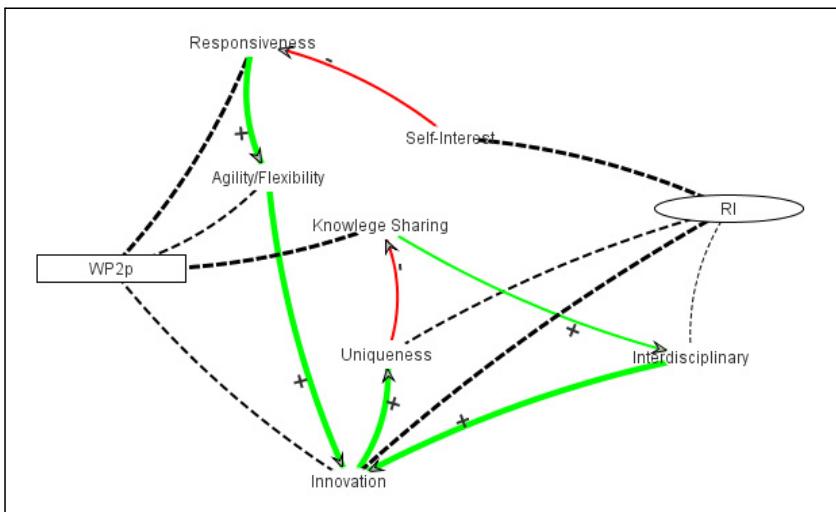


Fig. 3. Complete Aggregate Map for Research Center

Applying the qualitative assessment inference methods (described in detail in [7]) to evaluate the degree of alignment in terms of co-innovation characteristics of each innovation partner , the results presented in Table 1 can be achieved. Observing these results, it can be noticed that the Research Institute is not aligned with the Innovation Group's collaborative innovation evaluation perspective.

Furthermore, if we determine the *collaborative innovation values profile* (see Table 2) for each potential partner, we realize that group manager should pay special attention to the Research Institute behavior, since it presents some risk. This risk comes due to the fact that it defends *Uniqueness* (being unique) as an important characteristic, and the *Uniqueness* core-value has a negative influence on Knowledge Sharing, which is a core-value belonging to the collaborative innovation evaluation perspective, as it is illustrated in the complete aggregate core-values map presented in Fig. 3. Moreover, since this Research Institute is also characterized for acting according to its own interests (*Self-interest*), it is expected that it often does not be as responsive (*Responsiveness*) as required.

Table 2. Collaborative Innovation Values Profile

	Shared Values	Positive Impact	Negative Impact	Innovation Collaborative profile
Univ A	(innovation,high), (knowledgeSharing,very_high)	(innovation,strong), (flexibility,strong)		(innovation,very_high), (knowledgeSharing,very_high) (flexibility, high)
Univ B	(innovation,high), (knowledgeSharing,very_high)	(knowledgeSharing,moderate)		(innovation,high), (knowledgeSharing,very_high)
Research Institute	(innovation,high)		(uniqueness,high) (self – interest, moderate)	(innovation,high), (uniqueness, negative-high) (self – interest, moderate)
SME Network	(agility,high), (responsiveness,very_high)	(responsiveness,moderate)		(agility,high), (responsiveness,very_high)

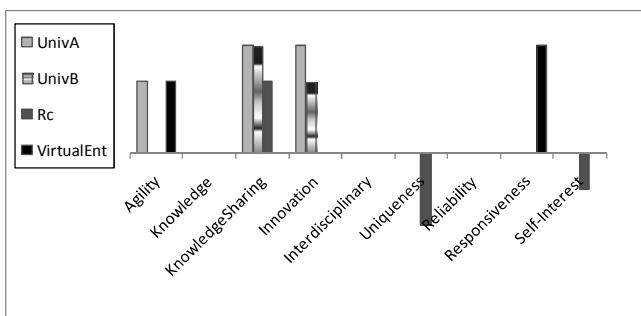


Fig. 4. Innovation Collaborative Values Profiles represented by a bar chart

Picking up the results presented in Table 2 a bar chart can be drawn (see Fig. 4) to visually illustrate the *collaborative innovation values profile* for each partner in the scope of this specific innovation opportunity. This chart evidences that the Research Institute ‘s *collaborative innovation values profile* has a negative impact on the collaboration opportunity. Moreover, we can easily notice that *Knowledge Sharing* is a relevant characteristic that contributes to the sustainability of this collaboration

relationship. *Responsiveness*, *Agility* and *Innovation* are also core-values that contribute positively to be aligned with the collaborative innovation values specified for this co-innovation opportunity.

6 Validation Process

The constructive research method [15] was the methodology selected to guide this research process. Following the constructive approach, a concept introduced through previous research can be applied to solve a specific problem, usually through the development of an artifact or a set of artifacts (models, diagrams, frameworks). March and Smith [16] claim that in this case, “the research contribution lies in the novelty of the artifact and in the persuasiveness of the claims that it is effective. Therefore, in order to validate the subsequent solution, two points have to be demonstrated: (i) that the artifact or set of artifacts proposed solve the domain problem and/or create knowledge about how the problem can be solved; (ii) how the solution proposed is new or better than previous ones.

The proposed Value System Management Model shows, to some extent, how it creates knowledge about how the problem can be solved. Moreover, the example presented above shows how the proposed approach solves the domain problem. The consistency of the CVS management model can be claimed, in view of the fact that it applies a set of models and methods previously verified and validated. Additionally, the novelty of the proposed approach can be claimed, because it was a first attempt to apply a Value System Management Model to manage value profiles in order to identify partners that are more adequate to collaborate in innovation process. However, the process of collecting evidences that allow us to claim the general usefulness of the design models and proposed method have to proceed. Thus, the development of a complete case study in the scope of the innovation networks is a future priority step towards that goal.

7 Conclusions and Future Work

This paper discussed how the implementation of a Core Value Management System can be useful for the establishment and sustainability of partnerships created following the co-innovation paradigm. Departing from a set of previously proposed tools, to specify and analyse Core Value Systems in collaborative environments, it has been described how they can be configured to determine collaborative innovation values profiles; and to assess if a potential partner has a values profile that is aligned with the partnership in terms of the core-values needed to work in such innovation network.

This approach has the following advantages: (i) it facilitates the representation of knowledge about the relevant core-values to the partnership in terms of innovation; (ii) it increases the “transparency” about the decision-making processes in the partner selection; (iii) it provides a visual representation of the interaction among partners in terms of core-values, allowing a better communication with the stakeholders.

Despite the work done so far suggests that the artefacts presented have a practical and theoretical relevance, the validation process for its complexity, has not yet been completed. Thus, the implementation of the CVS management model in a co-innovation network will be the next step.

Acknowledgements. This work was supported in part by the Portuguese “Fundação para a Ciência e a Tecnologia” (Portuguese Foundation for Science and Technology) through a PhD. scholarship.

References

1. Gloor, P.: *Swarm creativity: competitive advantage through collaborative innovation networks*. Oxford University Press, USA (2006)
2. Flores, M., Al-Ashaab, A., Magyar, A.: A Balanced Scorecard for Open Innovation: Measuring the Impact of Industry-University Collaboration. In: Camarinha-Matos, L.M., Paraskakis, I., Afsarmanesh, H. (eds.) PRO-VE 2009. IFIP AICT, vol. 307, pp. 23–32. Springer, Heidelberg (2009)
3. Chesbrough, H.: The era of open innovation. *Managing innovation and change* 127 (2006)
4. Macedo, P., Sapateiro, C., Filipe, J.: Distinct Approaches to Value Systems in Collaborative Networks Environments. In: Camarinha-Matos, L., Afsarmanesh, H., Ollus, M. (eds.) *Network-Centric Collaboration and Supporting Frameworks*, vol. 224, pp. 111–120. Springer, Boston (2006)
5. Camarinha-Matos, L.M., Macedo, P.: A conceptual model of value systems in collaborative networks. *Journal of Intelligent Manufacturing*, 1–13 (2008)
6. Macedo, P., Abreu, A., Camarinha-Matos, L.M.: A method to analyse the alignment of core values in collaborative networked organisations. *Production Planning & Control* 21, 145–159 (2010)
7. Macedo, P., Camarinha-Matos, L.M.: Applying Causal Reasoning to Analyze Value Systems. In: Camarinha-Matos, L.M., Pereira, P., Ribeiro, L. (eds.) DoCEIS 2010. IFIP AICT, vol. 314, pp. 3–13. Springer, Heidelberg (2010)
8. Rokeach, M.: *The nature of human values*. Free Press, New York (1973)
9. Schwartz, S.H.: Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. *Advances in Experimental Social Psychology*, 1–65 (1992)
10. Porter, M.: *Competitive Advantage*. The Free Press, New York (1985)
11. Alle, V.: Reconfiguring the Value Network. *Journal of Business Strategy* 21, 36–39 (2000)
12. Zineldin, M.A.: Towards an ecological collaborative relationship management A “co-opetive” perspective. *European Journal of Marketing* 32, 1138–1164 (1998)
13. Abreu, A., Camarinha-Matos, L.M.: On the Role of Value Systems and Reciprocity in Collaborative Environments. In: Spring (ed.): *Network-Centric Collaboration and Supporting Frameworks*. IFIP, vol. 224, Springer, Boston (2006)
14. Camarinha-Matos, L.M., Afsarmanesh, H.: ECOLEAD: A holistic approach to creation and management of dynamic virtual organizations. In: *Collaborative Networks and their Breeding Environments*, pp. 3–16. Springer, Valencia (2005)
15. Kasanen, E., Lukka, K., Siitonen, A.: The Constructive Approach in Management Accounting Research. *Journal of Management Accounting Research* 5, 245–266 (1993)
16. March, S.T., Smith, G.F.: Design and natural science research on information technology. *Decision Support Systems* 15, 251–266 (1995)

ProActive Service Entity Framework: Improving Service Selection Chances within Large Senior Professional Virtual Community Scenario

Tiago Cardoso and Luis M. Camarinha-Matos

Faculty of Science and Technology, Universidade Nova de Lisboa,
Quinta da Torre, Caparica, Portugal
`{tomfc, cam}@uninova.pt`

Abstract. Within a Collaborative Business Ecosystem context, as the network evolves, the competition between members increases and, as a consequence, members face the challenge to improve the chances their services have to be selected. On the other hand, the distance between Business Services and “Computational Services” is still a bottleneck for the automation of service provision. This paper extends the Pro-Active Service Entity Framework, proposing a mechanism to improve the service selection chances, based on the refinement of a Quality of Service concept for this context, and introduces a new class of actors for the framework – the intermediaries, needed to shorten the distance between Business and Computational perspectives.

Keywords: Business Services, Web-Services, Professional Virtual Community.

1 Introduction

Technological Services, such as Web Services, and Business Services (BSs), have gained the attention of the research community in the last decades, especially in the computer science area. The creation of the Web Services (WSs) approach, in a first stage, and the appearance of the Service Oriented Architectures, in a later stage, constitute nowadays the most commonly accepted and adopted mechanisms to support the development of ICT systems that support BSs. Nevertheless, although the advances in this technological perspective have significantly changed the way ICT systems support BSs there still is a gap between the way business people see the services they, or their enterprises, are willing to provide to customers or clients and the counterpart provided by the technological persons that develop the underlying ICT systems to support such provision. In other words, the business perspective of the services that either professionals or enterprises provide to clients is related with the client satisfaction, the service value or the resource management and the involved processes in such service provision. On the other hand, the ICT people’s perspective frequently focuses on interoperability, remote procedure calling or web-service

composition. The same gap exists within a Collaborative Business Ecosystem, where the members have a distinct way of seeing the services they are willing to provide to the community and the adopted web-services technological approach.

The passiveness of web services, in the sense they stay still waiting for the client's initiative, and the inexistence of aggregation between distinct Web-Services from the same provider are considered the major problems of current approaches, according to [1], where the authors propose a solution based on Multi-Agent Systems to target the passiveness problem, extending the Service Entity concept, first proposed in [2], in order to tackle the aggregation of distinct services provided by the same provider. Later, in [3] a first proposal is made for the Pro-Active Service Entity Framework (PASEF), extended in this paper, in order to create a mapping between Business and Software perspectives of service. In fact, PASEF constitutes a form of a Service Park, as identified in [4] bringing elements from the Multi-Agent Systems to the Collaborative Business Ecosystems, as foreseen in [5].

In this paper, the typical usage of PASEF within a Professional Virtual Community (PVC) of Senior Professionals is presented and the introduction of an additional class of actors is made – the PVC Intermediary. This actor has the role of bridging between the Senior PVC members and the clients, forming a team of value co-creation. One of the major tasks for this co-creation is the transformation of the business needs specified by clients into the identification and composition of the needed business services provided by the members of the PVC – the Senior Professionals. Finally, a mechanism for the calculation of Quality of Service is also proposed to support this task, especially useful in PVCs with a large number of members. This mechanism is inspired in [6], where bid evaluation considers client satisfaction, and [7], where bid arrival patterns are studied.

2 Contribution to Technological Innovation for Sustainability

Demographic evolution along with medical improvements that lead to longer average life expectation are two factors that compromise society's Sustainability. In other words, the number of working persons related to the number of retired persons are one of the major factors contributing for the world's Sustainability problems nowadays, as explained in [8]. Nevertheless, current trends point out that more people want to keep their active life after retirement and organizations are starting to support such "after retirement activity".

The contribution of this paper to Technological Innovation for Sustainability is aligned with these trends, proposing mechanisms to make Business and Software perspectives come closer within a senior PVC context. This proposal is based on the concept of value co-creation, evolving PVC Clients, the Senior Professionals and PVC Intermediaries, and based in a new mechanism proposed to assess Quality of Service from the Senior Professionals.

3 PASEF Proposal

The ProActive Service Entity Framework is intended to provide a solution within a competitive Collaborative Networked Business Ecosystem, particularly in a PVC, for the members of these communities to benefit from a pro-active representation of the Business Services they are willing to provide, as well as for clients to benefit from a wide range of potential Business Service providers.

The PASEF illustrating scenario is a Senior Coaching Association (SCA), a not for profit organization, dedicated to Senior Professionals that intend to continue their professional life after retirement, helping economic development through consultancy / coaching services, based on their professional life experience. This choice was made given the challenging demographic sustainability, as described in [8].

The framework is based on the assumption of the existence of a third class of actors, other than the PVC members and clients – the PVC Intermediary – with the role of facilitating the interface between the seniors and the business clients, namely through identification and characterization of business opportunities, as well as creation of Business Process Models based on the high-level needs specified by clients.

PASEF is composed of two complementary systems:

- Pro-Active Service Entity (PSE) – which aim to represent in a pro-active manner the Business Services a PVC member is willing to provide.
- PVC-Portal – a kind of PVC management system and virtual collaboration space that interacts with Clients who post their needs in high-level specifications and also the PVC Intermediaries and the PSEs.

The typical PASEF usage is divided in three stages: 1 - the PVC Member Registration stage, 2 - the Business Opportunity specification stage, and 3 - the Business Process Execution stage.

In the PVC Member Registration stage, each community member creates his / her PSE in the Portal and configures it, specifying the Business Services he / she is willing to provide, as well as the general availability and the

Name: Pedro; Email: pedro@sca.org , MAX BOs: Not defined MAX Simultaneous BOs: 5, BID Posting Policy: Auto BO Check rate: Week, Availability: Mon / Wed / Fri - (9:00 - 12:00)
--

(#, name, category, description, input info, result)

(1, "Live Marketing Strategy Definition", "Marketing Consultancy", "Marketing Consultancy with live meetings", "1st meeting date/time proposal", "PDF Report")
 (2, "Remote Marketing Strategy Definition", "Marketing Consultancy", "marketing consultancy withOUT live meetings", "documentation ZIP", "PDF Report")
 (3, "Marketing Results Monitoring", "Marketing Consultancy", "monitor marketing initiatives", "", "PDF Report")

Text Box 1. PSE Configuration example

PSE pro-activeness properties. Text Box 1 shows an example of a Senior that is willing to provide 3 consultancy services from the Marketing Consultancy Business Service Category. The first service, for example, is "Live Marketing Strategy Design", corresponding to a service that will be provided in some iterations and the input

information is the proposal for the first iteration date and time. The final result of this service is a PDF Report, corresponding to the Marketing Strategy.

The registration stage ends when the PSE is launched and starts looking for Business Opportunities where the services it represents are needed. After this stage, the PASEF manages a pool of services with pro-active behavior, in whose collaboration space clients can post their business needs and wait for the best proposals, based on an assessment of the Quality of Service made by PASEF, and other possible attributes, as explained below. In this Business Opportunity specification stage, the typical sequence of actions is summarized as follows:

1st – Clients make high-level specifications of their needs – the Business Opportunities. Textbox 2 shows a simplified illustration of a need specification.

2nd – PVC Intermediaries, in collaboration with the clients, transform such high-level specifications into abstract Business Process Models (absBPM), composed of Business Services from the PASEF pool. After the commitment of the Client with this absBPM, this step is concluded by posting a Call for Proposals (CfP), corresponding to each Business Service included in the absBPM, in a blackboard like interface within PASEF. Text box 3 shows a simplified CfP example from the same illustrative scenario, where the PVC Intermediary and the client decided that the needed absBPM for a Marketing Policy Coaching is the sequence of the Business Services: Marketing Strategy Definition, Marketing Plan Execution and Marketing Results Monitoring, as represented in Figure 1.

BO_ID - 2010-02-32
Company - FreshAir
Contact Person - José
Need Description: Marketing policy coaching

Text Box 2. Client Need

BO_ID - 2010-02-32
BID / Proposal Acceptance period - 5/2 - 15/2
Workflow Definition
[Order, Category ID, BS ID, name, input parameters, output]
[1, 77, 234, "Live Marketing Strategy Definition", [Meeting Date Proposal(Date/Time), documentation (ZIP)], [Report (PDF/Doc)]]
[2, 77, 456, "Marketing Plan Execution", [Marketing Plan (PDF/DOC)], [Final report (PDF/Doc)]]
[3, 77, 123, "Marketing Results Monitoring", [], [Final report (PDF/Doc)]]

Text Box 3. Client Need

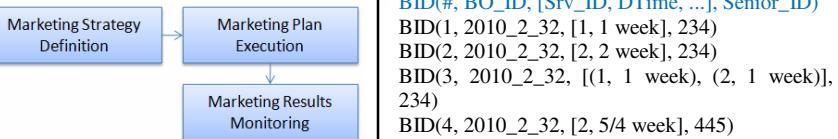


Fig. 1. Illustrative Workflow

BID(#, BO_ID, [Srv_ID, DTime, ...], Senior_ID)
BID(1, 2010_2_32, [1, 1 week], 234)
BID(2, 2010_2_32, [2, 2 week], 234)
BID(3, 2010_2_32, [(1, 1 week), (2, 1 week)], 234)
BID(4, 2010_2_32, [2, 5/4 week], 445)

Text Box 4. Illustrative Proposals

3rd – All PSEs look up for a suitable CfP, in a pre-defined frequency rate, to find opportunities where one or more of the services they represent are in need. Whenever this search has a successful result, they prepare / submit a proposal for the provision of such Business Service. Text box 4 shows an example of the posted proposals for the illustrative scenario. This is a special case where one PSE (ID: 234) represents two of the needed Business Services for a specific Business Opportunity. As a result

of this fact, the PSE prepares three proposals: one for each BS independently and one for the case of the provision of the two services.

4th – After a pre-defined time period or the arrival of a specific number of proposals, the PVC Intermediary closes the CfP and selects the best ones for each case, based on Quality of Service and other criteria, completing the executable BPM (eBPM). This step is concluded with another commitment of the Client, this time with this eBPM, corresponding to an agreement with the service providers' selection.

At a high level, this process can be seen as an instantiation of the well known contract-net protocol of the multi-agent systems area [9]. There are, however, several differences resulting from the combination of pro-activeness with the service notion.

Finally, after the eBPM is ready, with all the Business Services identified, as well as the Service Providers selected, the Business Process Execution stage starts when the Client launches the eBPM execution. From that point on, PASEF is responsible to call the included Business Services through the corresponding PSEs at the right time. The Service Providers are then informed to start the provision of their Business Services, eventually receiving needed input information in order to produce the corresponding output results.

One of the major objectives of this proposal is to bridge between the Business Services that PVC members provide and the technological services, Web-Services or other mechanisms, used to support this provision. In this context a Business Service is considered as an independent executing entity whose internal details, like task workflow or local data, are not taken into account, hence a BS is treated as a black box, although it must have an execution state and eventually input data and output results specification. Five distinct execution states are considered in a BS lifecycle:

1. Launch – the BS is ready to be launched by PASEF, when the time within the corresponding eBPM comes.
2. Ready – the PSE is ready to ask the community member it represents to start / continue the service provision / execution. This state happens in two situations: at the beginning of the BS lifecycle, when the service is launched, and whenever something the BS is waiting for becomes available - for example some documentation provided by another BS.
3. Run – the community member is executing the service.
4. Wait – the community member waits for some information from another community member, for example.
5. Complete – the BS gets complete and the results are eventually sent to PASEF, through the PSE, so they can be used as input parameters for subsequent BSs from the corresponding eBPM.

PASEF provides a toolbox for the PVC actors to trigger execution state transitions on such BSs, as well as to support any interaction between the PSEs of distinct BSs. Figure 2 represents the relation between the BS and PASEF Web-Services, for the case of a single BS execution. This execution state machine is repeated for each BS included in the eBPM.

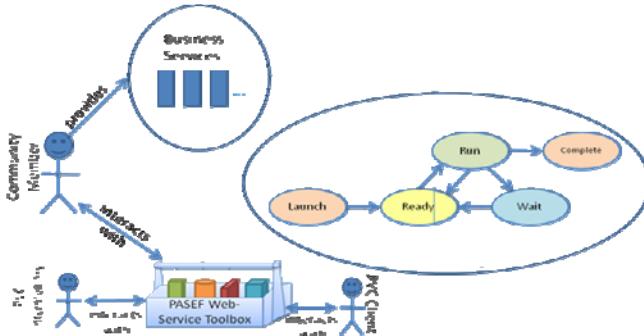


Fig. 2. PASEF bridge between Business and Technological Services

It is interesting to notice the similarity between the 5 identified BS states and the states a task can have in real-time-systems. Although the abstraction level is distinct, they both are execution entities, take their time executing and depend / interact with the surrounding environment. In the case of tasks from a real-time-system, the work is done by processors and the state-transitions-machinery is carried out by a Real-Time-Kernel. In the Case of these Business Services, the work is done by the PVC members and the state-transition-machinery is carried out through the PASEF toolbox.

3.1 PVC-Park and Pro-active Service Entity Web-Service functionality

The Web-Services provided by both PASEF PSEs and Portal are represented in Figure 3 and detailed in table 1 and 2.

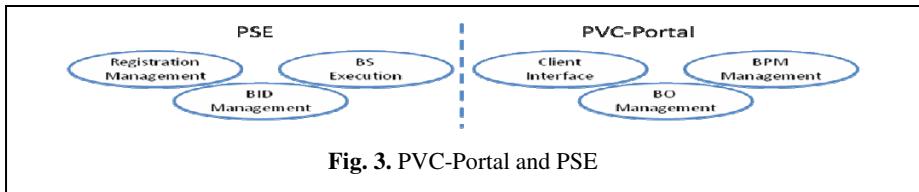


Fig. 3. PVC-Portal and PSE

Table 1. PASEF PSE WS Functionality

WebService	Functionality	WS Func.Client
Registration Managment	Service Registration	Member
BID Managment	BID Commit	Member
	BID Acceptance Notification	PVC Portal
Business Service Execution	Business Service Start	PVC Portal
	Business Service Cancel	PVC Portal

These Web-Services act as a toolbox for the three actors that interact with PASEF to be able to trigger the state transitions identified in Figure 2, in

Table 2. PASEF Portal WS Functionality

WebService	Functionality	WS Func.Client
Client Interface	New BO / New Need	CInt
	BO Search	PSE
	BID Submission	PSE
	BID Submit Acceptance Result	PVC_Interm.
BO Managment (Edition stage)	BPM Specification Submission	PVC_Interm.
	BPM Commit	CInt
BPM Managment (Execution stage)	eBPM Launch Execution	CInt
	BPM Cancel Execution	CInt
	BS Wait for Input	PSE
	BS End Notification / [result submission]	PSE
	BPM Execution Status	PVC Portal

order to make the bridge between Business Services from PVC members and this technological service set, in this case Web-Services.

3.2 PASEF Quality of Service

The mechanism used within PASEF to calculate the Quality of Service data is based on tracking all the service provision processes. Along time, PASEF stores information concerning all services, all service providers and all service provision instances. Based upon this information a configurable mechanism is provided for the PVC Intermediary, along with the Clients, to define the best way to classify and sort service provision proposals in order to make the best selection in all Business Opportunities. This mechanism is based in two concepts:

- QoSCharacteristic – Some property concerning the service that may be measured and compared between distinct services. This is the atomic mechanism information that is defined in terms of a name, the information that may be measured, and the unit used in that measurement. Two QoS Characteristic categories were identified concerning the:
 1. Service Provider - historic statistical information from the provider, e.g.:
 - N_BOs - number of BOs where the PVC member was involved,
 - BID_Success_Rate - BID Acceptance / Success Rate
 - Satisfaction - customer satisfaction level, graded in an after-service form,
 - OnTimeDeliverRate - % of on time delivery instances,
 - DelayOnDeliveryAverage - Average time of delay,
 - ...
 2. Service Provision - information from the Service Provision Proposal itself, specific to each proposal, e.g.:
 - DTime - proposed result delivery time,
 - ...
- QoS Criteria: the definition of the QoS Criteria consists in five tasks that have to be carried out by the PVC Intermediary with the help of the Client:
 1. Selection of the relevant QoS Characteristics for the specific Business Opportunity, e.g. N_BOs , DTime, Satisfaction.
 2. Definition / Selection of the evaluation schema, e.g., three classification levels: level 1(best), level 2 and level 3 (worst).
 3. Definition of how the selected QoS Characteristics “fit” into the specified schema, e.g.
 - N_BOs – Level 3 (values < 5); Level 1 (values > 15); Level 2 (otherwise)
 - DTime – Level 3 (values < 5); Level 1 (values > 15); Level 2 (otherwise)
 - Satisfaction – Level 3 (values < 5); Level 1 (values > 8); Level 2 (otherwise)

QoSCharacteristics Category weigh:
 Criteria schema: Level 1 / Level 2 / Leve 3.
 Restrictions:
 -Level 1 level mandatory for OnTimeDeliver
 -DTime < 15 and DelayOnDeliveryAverage < 4
 -Satisfaction not equal Level 3
 Final Grading Formula:

$$\frac{\text{OnTimeDeliver} * \text{Satisfaction}}{\text{N_Consortium_Members} * \text{N_BOs} * (15 - \text{DTime})}$$

Text Box 5. Sample QoS specification

4. Definition of restrictions, e.g. no proposals should be considered from providers with Level 3 classification concerning NBOs.
5. Definition of the overall rating formula for a provision proposal. This formula is created using the QoS Characteristics selected in 1 and is then used to classify all proposals to sort them in order to make the best selection.

4 Concluding Remarks

The extension of the PASEF proposal presented in this paper targets a better contribution for shortening the conceptual distance between business and software perspectives to the service concept. In fact, the adoption of a team built up by clients, providers and an intermediary in value co-creation brings to PASEF the missing roles a middle entity has to perform.

The application of this framework to the challenging area of active ageing is both actual and vital in the economic perspective, given the continuous increase in aged population and the needed demographic sustainability, among other factors.

Finally, with the definition of the mechanism for the assessment of Quality of Services, the PVC members see the excellence of their services rewarded.

As future work, the PVC Intermediary role has to be extended to the execution of eBPMs stage, in order to follow and track the evolution of the evolved processes and, eventually react to deviations on the pre-defined model. As a result, the corresponding ICT evolution will originate another extension of PASEF.

References

- [1] Cardoso, T., Camarinha-Matos, L.M.: Pro-active asset entities in collaborative networks. In: Camarinha-Matos, L.M., Pereira, P., Ribeiro, L. (eds.) DoCEIS 2010. IFIP AICT, vol. 314, pp. 93–102. Springer, Heidelberg (2010)
- [2] Franco, R.D., Bas, Á.O., Esteban, F.L.: Modeling extended manufacturing processes with service oriented entities. *Service Business* 3, 31–50 (2008)
- [3] Cardoso, T., Camarinha-Matos, L.M.: Pro-Active Service Entity Framework - Towards better mapping between Business and Software. In: Camarinha-Matos, L.M., Boucher, X., Afsarmanesh, H. (eds.) PRO-VE 2010. IFIP AICT, vol. 336, pp. 451–460. Springer, Heidelberg (2010)
- [4] Petrie, C., Bussler, C.: The Myth of Open Web-Services - The Rise of the Service Parks. *IEEE Internet Computing* (2008)
- [5] Camarinha-Matos, L.M.: Multi-Agent Systems in Virtual Enterprises. In: Proceedings of the International Conference on AI, Simulation and Planning in High Autonomy Systems (AIS 2002), SCS publication (2002)
- [6] Pang, N., Shi, Y., Ling, Y.: Bid Evaluation Method of Construction Project Based on Relevance and Relative Closeness. In: Proceedings of the International Conference on Information Management, Innovation Management and Industrial Engineering (2008)

- [7] Xing-jian, X., Shi, A., Bo, W.: Comparison Study of Duration and Bid Arrivals in eBay Auctions across Different Websites. In: Proceedings of the International Conference on Management Science & Engineering (2008)
- [8] Camarinha-Matos, L.M., Afsarmanesh, H.: Active Ageing Roadmap - A Collaborative Networks Contribution to Demographic Sustainability. In: Camarinha-Matos, L.M., Boucher, X., Afsarmanesh, H. (eds.) PRO-VE 2010. IFIP AICT, vol. 336, pp. 46–59. Springer, Heidelberg (2010)
- [9] Smith, R.G.: The Contract Net Protocol: High-level Communication and Control in a Distributed Problem. IEEE Transactions on Computers, 1104–1113 (1980)

Competence Mapping through Analysing Research Papers of a Scientific Community

Antonio P. Volpentesta and Alberto M. Felicetti

Department of Electronics, Computer Science and Systems,
University of Calabria
via P. Bucci, 42/C, 87036 Rende (CS), Italy
{volpentesta,afelicetti}@deis.unical.it

Abstract. Main goal of a scientific community is the collaborative production of new knowledge through research and scholarship. An integrative research approach, fostered by confrontation and collaboration among researchers, is widely recognized as a key factor to improve the quality of production of a scientific community. Competence mapping is a valid approach to highlight expertise, encourage re-use of knowledge, contributing significantly to the growth of the scientific community. In this paper we propose a methodological framework for examination and semi-manual classification of research papers. This method leads to the creation of a database that correlates research competences and researchers.

Keywords: Competence mapping, scientific community, collaborative knowledge, classification of research papers.

1 Introduction and Theoretical Background

Competence is a concept widely recognized in scientific literature since the early of 20th century, and particularly stressed in human resource management [1],[2]. Formally, competence is understood as the relation between humans and work tasks, specifically, which knowledge and skills are required to perform a specific task in an efficient way [3]. The concept of competence is related to the concept of competency; in [4] a competency is defined as “a specific, identifiable, definable, and measurable knowledge, skill, ability and/or other deployment-related characteristic (e.g. attitude, behavior, physical ability) which a human resource may possess and which is necessary for the performance of an activity”. In [5], in the effort of univocally identify HR concepts, authors define a competency as “the proven ability to use knowledge, skills and personal, social and/or methodological abilities, in work or study situations and in professional and personal development”, while a competence as “a competency (knowledge + skills + abilities) in a particular context (e.g. situation, domain)”. An effective competences management within an organization, i.e. oriented to the continuous enhancement and development of individual and organizational competences, requires as first and necessary step the mapping of organization competences. A Competences map can be defined a representation of key competences of each member of an organization [6]; it represents a valuable tool for identifying members of an organization who possess competences to perform a task. A competences map

facilitates efficient knowledge sharing between organizational members, constituting a multifaceted approach for creating structure out of an overabundance of potentially useful information [7],[8].

Traditional approaches in building competence maps are based on tools such as questionnaires or assessment sessions that allow organizations to define professional profiles of employees[9], [10]. In recent years, novel approaches based on document content analysis are taking place. Relational Content Analysis (RCA) approach, for instance, deals with extracting the relationships between actors, issues, values and facts from texts, for example relationships of support/criticism or cooperation/conflict between actors, subjective causal relationships between issue developments, or relationships between contents and actors [11]. Honkela et al. [12] claim that adaptive tools for data and text mining can be particularly useful in competence management enabling a more efficient process that can be based on a variety of heterogeneous sources of information.

Besides being important in the business oriented organizations, competence mapping is recognized to be important in many contexts. One of the most interesting, is the academic environment where competence mapping may provide a fruitful avenue for intellectual capital management [8]. In such context, and in particular within the scientific community focused on collaborative production of new knowledge, a problem currently faced is the inability of an organizations to know the competences owned by their members which prejudices the multi-disciplinary research and community creation [13].

A scientific community (also called epistemic community) is a collaborative network of professionals with recognized expertise and competence in a particular domain; such network defines policies for the collaborative creation of new knowledge within the target domain or issue-area [14]. Community members come from a variety of disciplines and backgrounds, having a shared set of normative and principled beliefs and a set of defined criteria for weighing and validating knowledge in the domain of expertise. As stated by Shur et al. [15], collaboration is the essence of science and is a key factor in scientific knowledge construction.

Unfortunately, collaboration in science is often reduced to the level of interpersonal knowledge. Scholars are not aware of other researchers who are working on similar projects thus collaboration frequently occurs among a small number of people working in the same group, dealing with or researching more specific items of their domain. A valid approach to highlight capabilities within a scientific community and suggest new collaborations between researchers may be the construction of a community competence map. This approach was already adopted by several authors. Rodrigues et al. [13] discuss about methods and techniques applied to the area of knowledge discovery from texts, in order to mapping researcher's competence in his/her publications. Janasik et al. [16] demonstrate the use of a self-organizing map method, a quantitative method for text mining, directed to organizational researchers interested in the use of qualitative data. In Vatanen et al. [17] a web environment is presented, called SOMPA (Self-Organizing Maps of Papers and Authors), for collecting and analyzing information on authors and their papers. None of these scholars faces up to the problem of defining research competences within a scientific community and to survey them by analysing the scientific production of community members. In this paper, the authors want to fill this gap proposing an original approach to map research competences

within a scientific community. The research approach makes use of a competence representation model, based on a logical structure of directed hypergraph [20], within a collaborative semi-manual mapping process.

2 Contribution to Sustainability

Usually the concept of sustainability is related to the use of a regenerative natural system in such a way that this system retains its essential properties [27]. However, in our concern, sustainability is defined as the infrastructure that remains in a scientific research community after the completion of research projects [28].

The goal of a scientific research community is to add new knowledge to that currently available in a particular field of inquiring, including the long-term maintenance of effects of new knowledge and fostering of collaboration between researchers. New scientific knowledge only becomes relevant to society, if it spills over. The concept of sustainability, in this terms, involves organizations that modify their actions as a result of participating in research, and individuals who, through the research process, gain knowledge and skills that are used in real world economic domains. Sustainability is more than the continuation of a research intervention, it concerns the possibility of maintaining or increasing effects achieved during a research phase. From this point of view, a more rational use of scientific competences certainly contribute to a more efficient approach in establishing a sustainability orientation of scientific research.

In this sense, research competences constitute an important component of sustainability as a part of the infrastructure that remains in a scientific research community after the completion of research projects.

An outcome of sustainability is the exchange of knowledge. Competence maps enable this exchange by making more effective and sustainable research collaborations and the development of new scientific questions and methods, [29].

3 Research Approach

Research is “an activity that aims to discover, interpret and revise in a non-trivial way, facts, events and theories, using a methodological approach in order to add new knowledge to that currently available, leading to new insights or improving the use of knowledge already available” [18]. This knowledge should be a wealth of structured information, easily accessible and of long-lasting value [19].

According to [5] definition of competency, we can define a *research competency* as “the proven ability to use knowledge, skills and personal, social and/or methodological abilities in conducting a research in a scientific domain”.

In this study, we use the terms :

- *Research competence (RC)*, as a “*research competency*” in a specific field of inquiring.
- *Researcher competence profile*, as the set of all research competences owned by a researcher.
- *Research competence map*, as a representation of a set of researchers competence profiles.

In what follows, a detailed description of an approach aimed to survey the research competence map for a selected scientific community is given. As stated before, the basic assumption for the approach is that it's possible to highlight the research competences of a researcher, and thus its competence profile, by analysing the researcher's scientific production (published papers). The approach is made of two fundamental parts: a competence representation model and a structured mapping process.

4 Competence Representation Model

Any scientific community is characterized by its own concept of research competence. Anyway, it is possible to affirm that generally a scientific community adopts a set of types of research and a scientific method on specific research fields of inquiring. This allows to affirm that a RC is strictly associated with a triple (t, p, s) , where:

- $t \in \mathcal{T}$, a set of Research Types;
- $p \in \mathcal{P}$, a set of phases of an inquiry method;
- $s \in \mathcal{S}$, a set of research subjects which are shared by the scientific community

In order to map competences in a scientific community, a competence representation model is necessary to be introduced. Such a model is used by community experts during the competence mapping process while performing a semi-manual analysis and classification of scientific production (research papers) published by members (researchers) of the selected scientific community. In what follows, we propose a model that is based on a logical structure of directed hypergraph [20] that represents relationships between concepts relating to research type (\mathcal{T}), phase of the scientific method (\mathcal{P}) and field of inquiry (\mathcal{S}).

Any node of the hypergraph fall into three categories:

- Undeveloped concept node
- Developed concept node
- Concept instantiation node

while any directed hyperedge represents how a concept is developed by connecting a set of nodes (tail of the hyperedge) with a developed node (head of the hyperedge). This means that the concept of RC may be represented by a directed hypergraph (H_0), consisting of a only one directed hyperedge as it is shown in fig.1:

The directed hypergraph representation highlights the relationship between the developed concept RC with the undeveloped concepts $\mathcal{T}, \mathcal{P}, \mathcal{S}$.

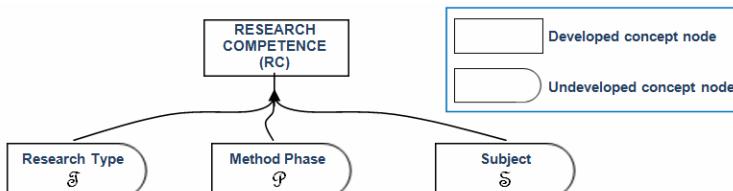


Fig. 1. Research Competence represented as directed hypergraph

One may recursively use this logical structure to represent the development of undeveloped concept nodes (in fig. 1, \mathcal{S} , \mathcal{P} , \mathcal{S}), by introducing concepts instantiation or new concepts to be further developed. In this way we extend the initial H_0 into a new one H_1 and so on. This process ends when, at a step n , any node that is not the head of any directed hyperedge in H_n is an instantiation node. Moreover, if we add a dummy node s and a simple hyperedge ($\{s\}$, x) for any instantiation node x , we obtain a (s,d)-hypernetwork [20]; an instantiation of the concept of RC remains associated to a directed hyperpath from s to d , where d is the RC node.

4.1 A Model Application: The Pro-VE Scientific Community Case

In order to illustrate the use of the model and the output of the representation process based on it, we may consider the Pro-VE Community, a scientific community that aims to promote research and production of new knowledge on Collaborative Networks [25].

This community adopts a scientific method which can be summarized as follows:

Research type means the type of contribution that is given to the body of available knowledge, the purpose of research and the methodological approaches used while researching. Based on these considerations, we identify the following categories of research:

- Theory (T): Research is focused on generation of new theories (theory building). This kind of research is generally based on an analytical approach that leads to the definition of a model (or a set of models) in order to give an interpretation of the variables involved, and offer predictions that can be verified. This research methodology comprises new insights by developing logical relationships between carefully defined concepts into an internally consistent theory [21]. It usually has low chance of success but, when successful, it has the potential to offer a high contribution to the wealth of knowledge.
- Empirical (E): Taking as a reference to the outcomes of purely theoretical research, research activities are focused on trying to draw general conclusions about the prospects of application, testing of existing theory and, marginally, extend theory. The empirical experimental research uses experimental design to verify the causality of a specific theory while elevating relationships from a testable hypothesis to an empirically verified theory. The outcomes of this kind of research will involve not only researchers and academics in general, but also companies who wish to assess the potential for transfer of results to their applications.
- Practices (P): research conducted with a practical and specific purpose, that aims to generate knowledge in the practical application of theories, developing insightful relationships between scientific research and real world application. It usually involves researchers and practitioners in a particular economic sector. This kind of research generally offers solutions to specific and practical problems, with an high chance of success but with a low contribution in new knowledge.

Phases of the inquiry method, in the case of Pro-VE community, are consistent with the scientific method phases. There are different ways of outlining the scientific

method used and procedures vary from one field of inquiry to another. However, components of the logical process of a scientific method can be broadly classified in 4 steps¹ that can be cyclically executed:

1. Characterizations: observation and description of a phenomenon, restating definitions in a new framework, and questioning what is known and what is unknown;
2. Hypotheses: theoretical, hypothetical explanations of observations or models of the subject inquiry;
3. Predictions: drawing consequences or making predictions from the hypothesis or theory, deriving results by means of a method;
4. Tests: conceptually or experimentally testing predictions or results and the path taken to them, measuring the usefulness of a model to explain, predict, and control, and of the cost of use of it.

By taking in consideration such concept developments, considering that the field of inquiring of Pro-VE community concerns the Collaborative Networks (CN), we obtain an extension of hypergraph in fig.1, as it is shown in fig. 2. We can identify a couple (x,y) where x is an instantiation of the concept of \mathfrak{T} , and y is an instantiation of the concept of \mathcal{P} , as a competency in a scientific community, while an instantiation of \mathfrak{S} is to be intended as a field of inquiry.

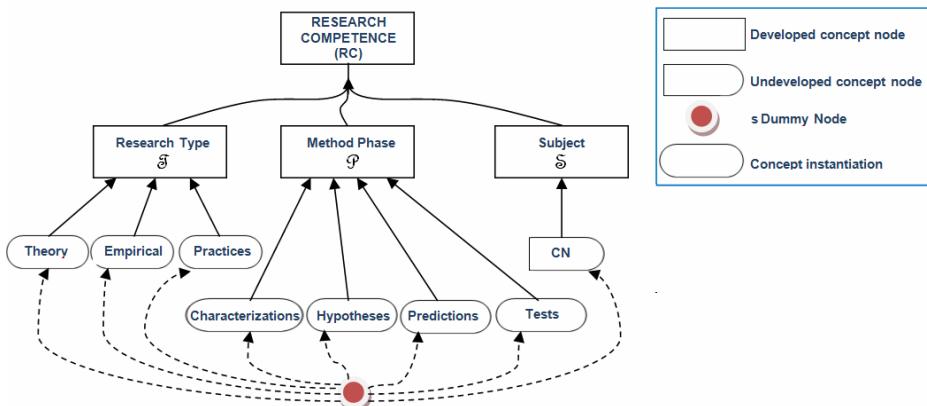


Fig. 2. H₁. First instantiation of the hypergraph in fig.1 to the case of the Pro-VE community.

Any hyperpath in the (s,d) -hypernetwork in Fig.2 is univocally identified by a triple (x,y,z) where (x,y) represents an instantiation of a scientific research competency and z is an instantiation of the concept of \mathfrak{S} .

To each pair (x,y) is associated a competency as showed in the following table:

A RC is a competency in a given field of inquiring. In order to show how RCs are represented in the model we may further develop the Hypergraph H₁. Next level H₂ can be obtained by developing the undeveloped node CN. In the field of CNs, a research

¹ This terminology is mainly used in natural sciences and social sciences. In mathematical sciences the terms “understanding”, “analysis”, “synthesis” and “review/extent” tend to be preferred in describing the method components, (Pólya,1957).

Table 1. Competencies Matrix for a generic scientific community

	Characterization	Hypotheses	Predictions	Tests
Practices	Finding and critically reviewing the background knowledge in a search for items that might help to put a theory and research results into practice.	Identifying a practice problem and formulating a general research question or model appropriate to this problem.	Providing innovative or exemplary practices, models or methods in communities, workplaces, organizations and the like.	Empirically validating and evaluating a model or a method in actual scenarios and examining its impact on current practices.
Empirical	Understanding an inquiry process of applying or testing a given theory under different perspectives or assumptions .	Generate research questions for a given theory by formulating research hypotheses (e.g. causal mechanism or a mathematical relation)	Using hypotheses to predict the existence of other phenomena, to predict quantitatively the results of new observations, or to draw some testable consequences.	Setting and performing empirical tests of an hypothesis or its consequences and evaluating them in the light of their compatibility with both the background knowledge and the fresh empirical evidence.
Theory	Reflecting upon one or more bodies of scientific literature or systems of thought and exploring the value of different theories and conceptual tools.	Providing theoretical evidence of important issues in identifying an explanatory gap in some theories.	Proposing and developing through logical reasoning a theoretical framework or concept in filling an epistemic gap.	Conceptually testing a theory, that is, checking whether results are compatible with the bulk of the existing knowledge on the matter.

topic concerns the study of a “Dimensional Aspects (concept substantially derived from the reference model described in [22]) of a CN Organizational Forms (derived from the classification provided in [23]; this study may relate or not an application in a economic sector in real business environments, where CN models, mechanisms, methodologies, principles and supporting tools are instantiated and implemented. Therefore the CN node can be developed as follows:

At this level, a RC is a competency in the study of a Dimensional Aspects of a CN Organizational Forms, or a competency in the study of a Dimensional Aspects of a CN Organizational Forms related to a particular economic sector.

Future developments of the Hypergraph lead to a more detailed representation of RCs.

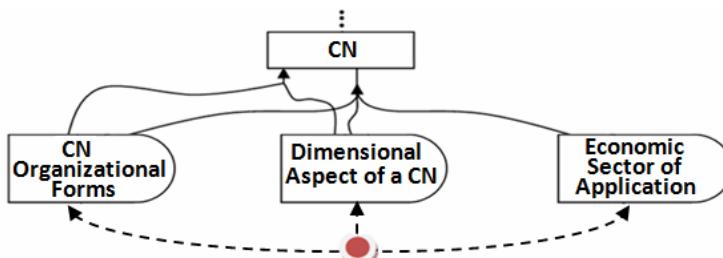


Fig. 3. Development of the research subject “Collaborative Networks” (CN) as represented in fig. 2

5 The Mapping Process

The mapping process is aimed to build a research competence map of a scientific community, that is to say a representation of relationships between research competences and researchers belonging to a scientific community.

This process uses the following inputs:

- Competencies Matrix for a generic scientific community.
- Initial Competence representation model for a scientific community (H_0)
- A list of researchers (authors).
- A set of scientific papers.

and provides the following outputs:

- A representation of relationships between research competences and researchers (community research competences map).
- A final Competence representation model for the scientific community (H_n).

The mapping process is based on the analysis of the scientific papers, performed by a team of expert. This methodological framework suggests that a steering committee, possibly formed by senior members of the scientific community who joined it since the beginning of its activities, is charged to define rules to both choose community experts (analyzers) and assign them papers to analyze.

The analysis of each scientific paper is based on three dimensions: the research approach adopted (Research Type), the phase of the scientific method (Method Phase) and the field of inquiry (Subject), according to Hypergraph competence model.

Supported by Competency Matrix, each expert have to individuate the competencies associated to the analyzed paper. In order to recognize the subject of the paper, each analyzer have to find opportune keywords associated to the concepts in the H_k competence model, through a semantic analysis of the papers' content. During the process, the competence model could be recursively updated to a more complete version.

This process allows to identify a set of RCs related to each paper. By assuming that research competences of any researcher are manifested on documents whose he is an author, is easy to associate to each researcher a list of RCs.

In order to map competences in a scientific community, we propose a model whose underlying logical structure is a vertex-hyperedge multi-hypergraph \mathcal{K} (R, \mathfrak{E}). The components of the model are:

- $R = \{r_1, \dots, r_m\}$ an ordered set of researchers (authors), members of a scientific community;
- $SC = \{sc_1, \dots, sc_n\}$ an ordered set of scientific research competences
- $A \in R^{m \times n}$ a binary matrix that represents the relationships between researcher r_i and research competence sc_k , i.e.: $a_{ik}=1$, if researcher r_i has the research competence sc_k , otherwise $a_{ik}=0$.
- $\mathfrak{E} = \{E_1, \dots, E_n\}$, with $E_j = E(sc_j) = \{ r_i \in R : k \text{ such that } a_{ik}=1 \}$. E_j is a subset of R consisting of all researchers that share the research competence sc_j . Of course, a researcher may have many research competences and many research competences may be shared by the same subset of researchers (this is the reason why \mathcal{K} is a multi-hypergraph).

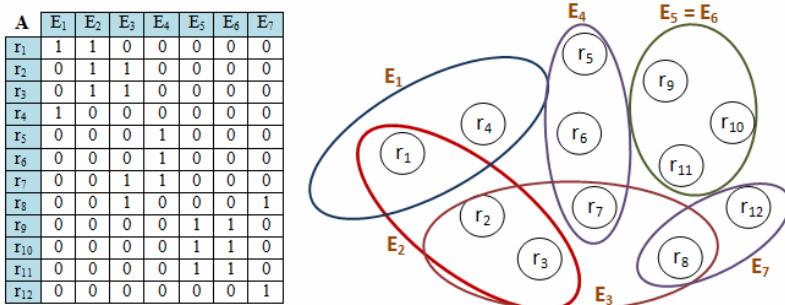


Fig. 4. Example of research competence map

In order to better clarify the mapping process, we present a representation based on 3 phases according to the IDEF0 notation [24]:

5.1 Phase 1: Activities Planning

- Actors and roles: a steering committee, possibly formed by senior members of the community who joined it since the beginning of its activities, is charged to define rules to both choose community experts (analyzers) and assign them papers to analyze, the rules for paper analysis, and a software tools that support the activity of analysis.
- Inputs: list of all researchers gathered around the community and, for each of them, the published papers;
- Controls: rules to choose analyzers, rules to assign papers to analyzers
- Mechanisms: steering committee.
- Outputs: list of analyzers, assignment of papers to analyzers, paper classification procedure;

5.2 Phase 2: Paper Analysis

- Actors and Roles: analyzers, charged to apply paper classification procedure in order to analyze the papers in the repository; steering committee, who control and validate outputs of the procedure
- Inputs: Competencies Matrix for a generic scientific community, H_0 , list of analyzers, assignment of papers to analyzers;
- Controls: paper classification procedure.
- Mechanisms: steering committee, analyzers, software tools for papers analysis, papers repository
- Outputs: definitive competence representation model (H_n), database of classified papers

The *paper classification procedure* is an algorithm that comprises n steps as follows:

Step k ($1 \leq k \leq n$): Supported by Competencies Matrix, each analyzer has to process its assigned papers in order to recognize scientific approaches and the phases of scientific

method present in each paper. Furthermore he extract from each article, the keywords related to the concepts described in H_{k-1} .

At the end of this step, each analyzer sends to the steering committee the results of his classification. The steering committee discuss and validate all the new keywords discovered by analyzers at this step and decide how to extend H_{k-1} to H_k . This extension consists in using the validated new keywords to develop undeveloped concept nodes, by introducing concepts instantiation or new concepts to be further developed.

Step n: When neither further keywords are provided by analyzers or new extensions are defined by the steering committee on H_{n-1} (meaning that any terminal node in H_{n-1} is an instantiation node), then the algorithm terminates and the procedure gives in output:

- $H_n \equiv H_{n-1}$, i.e. the definitive competence representation model for the community;
- the set of all papers classified according H_n , i.e. the database of classified papers;

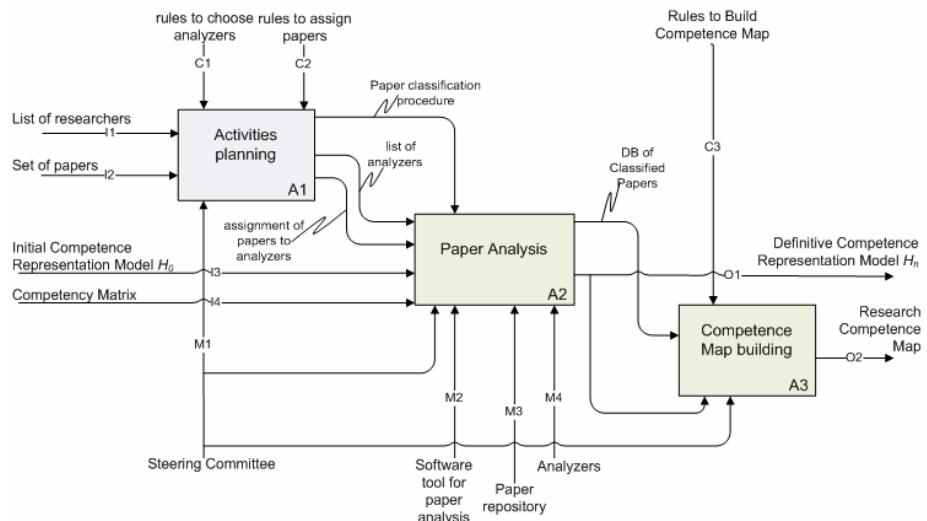


Fig. 5. Graphical representation in IDEF0 notation of the mapping process

5.3 Phase 3: Competence Map Building

- Actors and roles: the steering committee, analyzes the classification of scientific papers according to H_n in order to build a research competence map of the scientific community.
- Inputs: database of classified papers,
- Controls: rules to build competence map.
- Mechanisms: steering committee, competence representation model (H_n),
- Outputs: research competence map.

6 Conclusions and Future Works

The presented work is aimed to propose a methodological framework for examination and classification of research papers, useful to build a competence map of a scientific community, relying on the correlation between research competences and researchers. This methodological framework is based on a competence representation model Competence, that is based on a logical structure of directed Hypergraph, to represent the competences present within a scientific community.

In order to represent the competence map of a scientific community, we utilize a multi-hypergraph structure, that models the relationships between researchers and research competences.

An implementation of the proposed research approach will be carry out in order to build a competence map of the Pro-VE community, by examining and classifying the papers published in the last five books related to the Pro-VE Conferences (from 2005 to 2009) [26]. This leads to the creation of a database that correlates research competences and researchers, in order to perform statistical analysis of competences within the PRO-VE scientific community.

References

1. Taylor, F.: *The Principles of Scientific Management*. Harper & Row, New York (1911)
2. Boyatzis, R.E.: *The Competent Manager: A Model for Effective Performance*. John Wiley and Sons, New York (1982)
3. McLelland, D.C.: Testing for Competence rather than for "Intelligence". *American Psychologist* 28, 1–14 (1973)
4. Allen, C.: Competencies (Measurable Characteristics), Recommendation of the HR-XML consortium, February 26 (2003), http://ns.hr-xml.org/2_0/HR-XML-2_0/CPO/Competencies.pdf (March 2010)
5. Vervenne, L., Najjar, J., Ostyn, C.: Competency Data Management (CDM), a proposed reference model. European Commission – Semantic Interoperability Centre Europe (April 2010), <http://www.semic.eu/semic/view/documents/Competency-Related-Data-Management.pdf> (retrieved)
6. Draganidis, F., Mentzas, G.: Competency based management: a review of systems and approaches. *Information Management & Computer Security* 14(1), 51–64 (2006) ISSN: 0968-5227
7. Wexler, M.N.: The who, what and why of knowledge mapping. *Journal of Knowledge Management* 5(3) (2001)
8. Hellstrom, T., Husted, K.: Mapping knowledge and intellectual capital in academic environments: A focus group study. *Journal of Intellectual Capital* 5(1), 165–180 (2004)
9. Liang, Y., Konosu, T.: A Study on Competency Evaluation of the Members in Software Development Project. *Journal of the Society of Project Management* 6(2), 29–34 (2004)
10. Yonamine1, R.K., Nakao, O.S., Martini, J.S.C., Grimon, J.A.B.: A program for the professional development of Brazilian engineering students: origin and development. In: *Proceedings of the International Conference on Engineering Education & Research (ICEE & ICEER KOREA)* (2009)
11. Van Atteveldt, W., Kleinnijenhuis, J., Oegema, D., Schlobach, S.: Representing Social and Cognitive Networks. In: *Proceedings of the 2nd Workshop on Semantic Network Analysis*, Budva, Montenegro, June 12 (2006)

12. Honkela, T., Nordfors, R., Tuuli, R.: Document maps for competence management. In: Proceedings of the Symposium on Professional Practice in AI. IFIP, pp. 31–39 (2004)
13. Rodrigues, S., Oliveira, J., Moreira de Souza, J.: Competence mining for virtual scientific community creation. *Int. J. Web Based Communities* (2004)
14. Haas, P.M.: Epistemic Communities and International Policy Coordination, International Organization. *Knowledge, Power, and International Policy Coordination* 46(1), 1–35 (1992)
15. Schur, A., Keating, K.A., Payne, D.A., Valdez, T., Yates, K.R., Myers, J.D.: Collaborative suites for experiment-oriented scientific research. *Interactions* 3, 40–47 (1998)
16. Janasik, N., Honkela, T., Bruun, H.: Text Mining in Qualitative Research: Application of an Unsupervised Learning Method. *Organizational Research Methods* 12(3) (2009)
17. Vatanen, T., Paukkeri, M., Nieminen, I.T., Honkela, T.: Analyzing authors and articles using keyword extraction, self-organizing map and graph algorithms. In: Proceedings of the AKRR 2008, pp. 105–111 (2008)
18. Howard, K., Sharp, J.: The Management of a student research project (1983)
19. Melnyk, A., Handfield, R.: May you live in interesting times.. the emergence of theory-driven empirical research. *Journal of Operation Management* (1998)
20. Volpentesta, A.P.: Hypernetworks in a directed hypergraph. *European Journal of Operational Research* 188, 390–405 (2008)
21. Wacker, J.G.: A definition of theory: research guidelines for different theory-building research methods in operations management. *Journal of Operations Management* 16 (1998)
22. Romero, D., Galeano, N., Molina, A.: A Virtual Breeding Environment reference model and its instantiation methodology. In: Camarinha-Matos, L.M., Picard, W. (eds.) *Pervasive Collaborative Networks*. Springer, Boston (2008)
23. Camarinha-Matos, L.M., Afsarmanesh, H.: Collaborative Networks: Value Creation in a Knowledge Society. In: *Knowledge Enterprise*. IFIP, vol. 207, pp. 26–40. Springer, Boston (2006)
24. Jeong, K.Y., Wu, L., Hong, J.D.: IDEF method-based simulation model design and development. *Journal of Industrial Engineering and Management* 2(2) (2009)
25. <http://www.pro-ve.org>
26. Camarinha-Matos, L.M., et al. (eds.): *Pro-VE 2005 IFIP Vol. 185, Pro-VE 2006 IFIP Vol. 224, Pro-VE 2007 IFIP Vol. 243, Pro-VE 2008 IFIP Vol. 283*, Springer, Boston, *Pro-VE 2009 IFIP Vol. 307*. Springer, Heidelberg
27. Schmoch, U., Schubert, T.: Sustainability of incentives for excellent research – The German case. *Scientometrics* 81(1), 195–218 (2009)
28. Altman, D.G.: Sustaining Interventions in Community Systems: On the Relationship Between Researchers and Communities. *Health Psychology* 14(6), 526–536 (1995)
29. Lee, Y.S.: The Sustainability of University-Industry Research Collaboration: An Empirical Assessment. *Journal of Technology Transfer* 25, 111–133 (2000)

Tuple-Based Semantic and Structural Mapping for a Sustainable Interoperability

Carlos Agostinho^{1,2}, João Sarraipa^{1,2}, David Goncalves¹,
and Ricardo Jardim-Goncalves^{1,2}

¹ Departamento de Engenharia Electrotécnica, Faculdade de Ciências e Tecnologia, FCT,
Universidade Nova de Lisboa, 2829-516 Caparica, Portugal

djg15361@fct.unl.pt

² Centre of Technology and Systems, CTS, UNINOVA, 2829-516 Caparica, Portugal
{ca,jfss,rg}@uninova.pt

Abstract. Enterprises are demanded to collaborate and establish partnerships to reach global business and markets. However, due to the different sources of models and semantics, organizations are experiencing difficulties exchanging vital information electronically and seamlessly, even when they operate in related business environments. This situation is even worst in the advent of the evolution of the enterprise systems and applications, whose dynamics result in increasing the interoperability problem due to the continuous need for model adjustments and semantics harmonization. To contribute for a long term stable interoperable enterprise operating environment, the authors propose the integration of traceability functionalities in information systems as a way to support such sustainability. Either data, semantic, and structural mappings between partner enterprises in the complex network should be modelled as tuples and stored in a knowledge base for communication support with reasoning capabilities, thus allowing to trace, monitor and support the stability maintenance of a system's interoperable state.

Keywords: Interoperability, Model Morphisms, Semantic Matching, Knowledge Representation, Sustainable Interoperability.

1 Introduction

In the emerging society, characterized by the globalization phenomena, technological evolution and constant financial fluctuations, knowledge is a major asset in people's lives. It is and will remain being, in the future, the principal factor for competition both at personal and organizational levels, conducting tendencies at global markets. The outburst of advanced Web technologies, knowledge bases and resources all over the world, is levelling markets as never, and enabling organizations to compete on an equal basis independently of their size and origin [1].

The traditional way of doing business is not providing the expected efficiency. Nowadays, companies do not survive and prosper solely through their own individual efforts. Each one's success also depends on the activities and performance of others to whom they do business with, and hence on the nature and quality of the direct and

indirect relations [2]. These involve a mix of cooperative and competitive elements, and to cope with them, organizations need to focus on their core competencies by improving their relationships with customers, streamlining their supply chains (SCs), and by collaborating with partners to create valued networks between buyers, vendors and suppliers [3][4][5]. Indeed, in most cases, a single company cannot satisfy all customers' requirements, and where once individual organizations battled against each other, today the war is waged between networks of interconnected organisations [6], e.g. SCs. Therefore, to succeed in this complex environment, enterprise systems need to be interoperable, thus being able to share technical and business information, within and across organisations in a seamless and sustainable manner [5][7]. In this sense, sustainability appears on the context of this paper, related to the information systems (IS) ability to smoothly accommodate technical disturbances in a network of organisations, without compromising the overall network interoperability state.

2 Contribution to Technological Innovation and Sustainability

If systems are only partially interoperable, translation or data re-entry is required for information flows, thus incurring on several types of costs. For example, in SCs if the lower tiers do not have the financial resources or technical capability to support interoperability, their internal processes and communications are likely to be significantly less efficient, thus harming the performance of the entire network. This way, achieving an interoperable state inside heterogeneous networks, is still an ongoing challenge hindered by the fact that they are, intrinsically, composed by many distributed hardware and software using different models and semantics [8]. This situation is even worst in the advent of the evolution of enterprise systems and applications, whose dynamics result in increasing the interoperability problem with the continuous need for model adjustments and semantics harmonization: retail and manufacturing systems are constantly adapting to new market and customer requirements, thus answering the need to respond with faster and better quality production; new organizations are constantly entering and leaving collaboration networks, leading to a constant fluctuation and evolution of system models. All these factors are making interoperability difficult to sustain [9].

Due to this constant knowledge change, ontologies and model mappings are not static and there is always some information to add to a knowledge representation system. Preliminary theories have been advanced in specific scientific disciplines, such as biology and ecology, to explain the importance and evolution of complexity in living systems [10]. Also, some researchers have attempted to extrapolate results from a "general systems theory" or "complexity theory" that could explain the importance of the behaviour of systems in all fields of science [11][12]. These theories view all systems as dynamic, "living" entities that are goal-oriented and evolve over time, thus, IS should be able to manage its dynamics, learning during its existence and being in a constant update [13].

Based on that assumption, this paper contributes for a sustainable interoperable environment, proposing an approach inspired on complex adaptive systems (CAS) that use monitoring and traceability functionalities to act at the network micro level (i.e. local IS) and support sustainability at the macro level (i.e. the network). Data, semantic, and structural mappings are proposed to be modelled as traceable tuples and

integrated in knowledge bases dedicated to managing mismatches during communications. Section 3 summarizes different ways to represent and formalize model morphisms and semantic matching; Section 4 defines the concept of tuple for semantic and structural mapping, as well as the knowledge base for communication support; Section 5 presents a case study scenario for validation; and finally, in Section 6, the authors conclude and outlook on future work.

3 Models and Associated Concepts

3.1 Models

Either being used in the form of traditional databases, architectural models, or domain ontologies, models can be described on multiple formats, languages, expressiveness levels, and for different purposes [14][15][16]. A model can be characterized according to four dimensions [17]: *Metamodel* - the modelling primitives of the language for modelling (e.g. ER, OWL, XSD) are represented by a set of labels defined in the metamodel; *Structure* - corresponding to the topology associated to the model schema; *Terminology* - the labels of the model elements that don't refer to modelling primitives; *Semantics* - given a "Universe of Discourse", the interpretations that can be associated with the model.

This way, model operations can be classified as acting on any of these dimensions.

3.2 Model Morphims (MoMo)

In mathematics, "Morphism" is an abstraction of a structure-preserving map between two mathematical structures. It can be seen as a function in set theory, or the connection between domain and co-domain in category theory [17]. Recently, this concept has been gaining momentum applied to computer science, namely to systems interoperability. This new usage of "morphism" specifies the relations (e.g. mapping, merging, transformation, etc) between two or more information model specifications (M as the set of models). Therefore, a MoMo describes a model operation.

In this context, the research community identifies two core classes of MoMo: non-altering and model altering morphisms [17][18]. As evidenced in Table 1, in the non-altering morphisms, given two models (source A and target B), a mapping is created relating each element of the source with a correspondent element in the target, leaving both models intact. In model altering morphisms, the source model is transformed using a function that applies a mapping to the source model and outputs the target model [19]. Other relations, such as the merge operation, can also be classified as model altering morphisms, however they are not detailed in this paper.

Being more interested in the mapping operation (non-altering) for this paper, these generic function descriptions are not detailed enough to deal with the specificities of the multiple information models used by the enterprise systems of today's business networks. To respond to the constant knowledge and model changes on heterogeneous and dynamic networks, it is required to use a more detailed and traceable mapping format that provides a semantic "link" between two different models and its components. On the following sub-sections, technologies and formalization methods will be analysed concerning their usability towards that goal.

Table 1. Cases of Model Morphisms

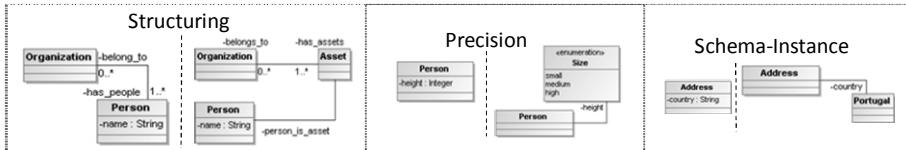
MoMo	Formalization	Classification
Mapping: $\theta(A, B)$	$\forall A, B \in M: \theta(A, B) \subseteq Sub(A) \times Sub(B)$	Non-altering
Transformation: $\tau: A \times \theta \rightarrow B$	$\forall A, B \in M: \text{if } \exists \theta(A, B) \text{ then } \tau(A, \theta) = B$	Model altering

3.3 Semantic Mismatches

Mismatches are inconsistencies of information that result from “imperfect” mappings. Due to the differences among models referred before, almost in every case, a MoMo leads to a semantic mismatch, which can either be lossy or lossless depending on the nature of the related model elements (see Table 2): In lossless cases, the relating element can fully capture the semantics of the related; while in lossy mismatches a semantic preserving mapping to the reference model cannot be built [20].

Table 2. Semantic Mismatches (based on [20])

Mismatch		Description
Lossless	Naming	Different labels for same concept
	Granularity	Same information decomposed (sub)attributes (see Figure 2)
	Structuring	Different design structures for same information (see Figure 1)
	SubClass-Attribute	An attribute, with a predefined value set (e.g. enumeration) represented by a subclass hierarchy
	Schema-Instance	An attribute value in one model can be a part of the other’s model schema (see Figure 1)
	Encoding	Different formats of data or units of measure (e.g. kg and lbs)
Lossy	Content	Different content denoted by the same concept
	Coverage	Absence of information
	Precision	Accuracy of information (see Figure 1)
	Abstraction	Level of specialisation (e.g. “Car” and “Ford”)

**Fig. 1.** Mismatch examples

This notion of mismatch can bring a semantic meaning to the type of the relationship being established in the mapping. However, the envisaged semantic “link” between two different models needs to account for more than inference of a meaning. It needs to be represented through a formal expression that is traceable and parseable by an intelligent system that can deduce and recommend mapping readjustments, which might even change the mismatch type.

3.4 MoMo Formalisms

Model Morphisms, as envisaged in section 3.2, are intended to introduce a method of describing relationships/transformations among models. Originally graph theory has been used, but other and theories can be considered to achieve the envisaged goals:

Classical Mathematics: Graph & Set Theory

Graphs are a common way to graphically present models, where the nodes are considered as a domain entity and the edges as relations between them. For the purposes of MoMo, model operations such as the ones of Table 1 can be described using a 6-tuple labelled oriented multigraph (*LDMGraph*) of the form $G=(V,E,s,t,lv,le)$, where: V is the vertex set of G ; E is the edge set of G ; $s: E \rightarrow V$, is a function that associates an edge with its source vertex; $t: E \rightarrow V$, is a function that associates an edge with its target vertex; $lv: V \rightarrow \sum V$, is a function that associates a vertex with its label; $le: E \rightarrow \sum E$, is a function that associates an edge with its label [17], [21]. This abstract view of models allows formal reasoning on their properties and on the properties of the model operations needed for their effective management.

As graphs, also sets can be used to represent models and operations using first-order logic, algebra and axioms. Being defined as a collection “ M ” of distinct objects “ m ”, a set can represent objects, numbers, other sets, etc [22]. Operations such as membership “ $M1 \subseteq M2$ ”, power “ $P(M)$ ”, union “ $M1 \cup M2$ ”, intersection “ $M1 \cap M2$ ”, complement “ $M1 \setminus M2$ ”, or cartesian product “ $M1 \times M2$ ” are already well defined.

Mapping as a model: Model Management [23]

This theory defends that a mapping between models $M1$ and $M2$ should be a model “ $map12$ ” and two morphisms (one between “ $map12$ ” and $M1$ and another between “ $map12$ ” and $M2$). Thus, each object “ m ” in the mapping can relate a set of objects in $M1$ to a set of objects in $M2$. In this approach, instead of representing a mapping as a pair of objects, a mapping is represented as a set of objects (see Figure 2). Using concepts from classical mathematics, this approach enables to define complex algebra to describe major model operations such as match, compose, diff, model gen, or merge.

Mapping as a complex tuple: Matching [24]

The match operator takes two graph-like structures and produces a mapping between the nodes of the graphs that correspond semantically to each other. Mappings between these elements can be described using set-theoretic semantic relations instead of using traditional numeric coefficients. The meaning of concepts (not labels) within a model can determine equivalence “ $=$ ”, more “ \sqsupseteq ” and less “ \sqsubseteq ” general, as well as disjointness “ \perp ” relationships. Having this, a mapping element can be defined as a 4 level tuple $<IDij, ai, bj, R>$ where: $IDij$ is a unique identifier of the given mapping element; ai is the i-th node (or vertex) of the first tree; bj is the j-th node of the second tree; and R specifies the semantic relation which may hold between them.

The above methodologies seem to be powerful in terms of expressiveness of the morphism. However others exist, such as the composition of complex operations

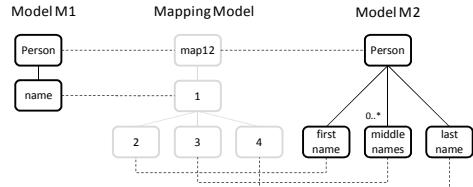


Fig. 2. Mapping as a model ($map12$)

based on a catalogue of primitive transformations [25]. However, this approach is more focused on model altering morphisms.

4 Modelling Morphisms to Enable Sustainable Interoperability

So far, a proven approach to deal with interoperability relies on the usage of dedicated knowledge models and international standards acting as information regulators among organizations. However, due a complexity of reasons many organizations are still focused on P2P relationships, where each one tends to use its own data format and business rules, and handles as many mappings as the number of business partners [9].

Either case, after interoperability is first established and all morphisms defined, the set of organizations within a network demonstrate a period of stability exchanging e-messages following the established mappings [9]. At this stage, the networks display symmetry [26]. However, that might not be sustainable for long if business requirements change. Organizations are managed by people that have different opinions and backgrounds based on several factors such as culture, professional experience, family, etc. They manage, work, and are themselves customers of different organizations, which in turn have different systems that are structured according to several information models implemented on multiple software platforms. All this heterogeneity leads in most cases, the network to experience problems because if just one of the network members adapts to a new requirement, the harmony is broken, and the network begins experiencing interoperability failure. This is even more evident in multi-domain networks (e.g. collaborative product design) where information is dispersed and frequently replicated in many IS.

To mitigate that, and inspired by CAS, context awareness and traceable morphisms are demanded in support of intelligence. Also, monitoring and decision support systems must be considered in the construction of a framework that implements sustainable interoperability in cooperation networks, thus addressing the robustness of an IS both at conceptual and structure related levels. The sustainable interoperability targeted in this paper is mostly technical with direct impacts on the economical axis of sustainability, but will also produce indirect effects on the social and environmental axis since it enables organizations to redirect money away from technological issues.

4.1 Knowledge Enriched Tuple for Mappings Representation

Observing all previously explained technologies and methodologies for managing morphisms, the authors consider that there is no perfect solution that can provide all the desired goals at once. Some are ideal for structural issues, others for semantics providing good human traceability, while others are more formal and mathematical based. Therefore, we propose the usage of a 5-tuple mapping expression (equation 1), reusing some of the concepts explained in section 3, that formalizes the morphism between two model elements (a and b) and is enriched with semantic information that enables fast human readability, where $\forall A, B \in M, \exists a \in A \text{ and } \exists b \in B: \text{if } M \text{ is an LDMGraph then } a \in V(A) \text{ and } b \in V(B)$.

$$\text{Mapping Tuple (MapT): } < ID, MElems, KMTtype, MatchClass, Exp > \quad (1)$$

- ID is the unique identifier of the MapT and can be directly associated with the a 's vertex number: $IDi.j_x: 1 \leq i \leq |V(A)| \text{ and } 1 \leq j \leq |V(Sub(B))| \text{ and } x \in \mathbb{N}$. The depth of the sub-graph detail used in the mapping is not limited, and x is a counter for multiple tuples associated with the same concept;
- $MElems$ is the pair (a, b) that indicates the mapped elements. If the ID specifies a mapping at the n -th depth level of the graph, a should be at the same level, i.e. $a.ai$ (for $i = 1..n$);
- $KMType$ stands for Knowledge Mapping Type, and can be classified as: “Conceptual” if mapping concepts and terms; “Semantics” if mapping model schemas; and “InstantiableData” if the mapping is specifying instantiation rules.
- $KMType = \{Conceptual, Semantics, InstantiableData\}$;
- $MatchClass$ stands for Match/Mismatch Classification and depends on $KMType$, such as $\forall(a, b) \in MElems$:
- $\forall KMType$, if $a=b$, the mapping is absolute and $MatchClass = Equal$;
 - if $KMType = Conceptual$, the mapping is relating terms/concepts, and $MatchClass \in \{Equal, Naming, Coverage, MoreGeneral, LessGeneral, Disjoint\}$ depending on the coverage of the relationship;
 - Otherwise, the mapping is structural or non-existent and $MatchClass \in \{Equal, Disjoint\}$;
- Exp stands for the mapping expression that translates and further specifies the previous tuple components. It can be written using a finite set of binary operators derived from the mathematical symbols associated with the mapping types and classes (e.g. “ $=, \sim, \subseteq, \supseteq, \perp, +, -, \times, \div, concatenate, split$ ”).

This mapping tuple which represents $\theta(a, b)$, can also be used to generate a transformation function τ , where $\tau(a, \theta) = b$, being $(a, b) \in MElems$. Therefore, when used by intelligent systems such as CAS-like IS, the tuple's information enables automatic data transformations and exchange between two organizations working with/on different information models, thus achieving an interoperable state among them and supporting the recovery from any harmonization breaking situation.

4.2 Communication Mediator (CM)

With the MapT stored in a knowledge base (KB) to support communications intelligence, all information concerning the mappings between models or ontologies of business partners can be accessed by their local systems. This allows communities to build IS with reasoning capabilities able to understand each others' representation format, without having to change their data and communication functions [13].

The proposed CM is defined by an ontology in OWL format. It has been built up as an extension to the Model Traceability Ontology defined in [27], which addresses traceability as the ability to chronologically interrelate the uniquely identifiable objects in a way that can be processed by a human or a system. The structure of the evolved communication mediator is presented in Figure 4 and described as follows.

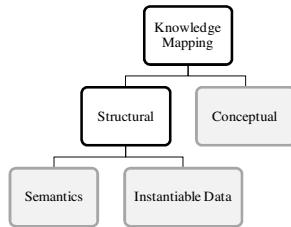


Fig. 3. KMType values

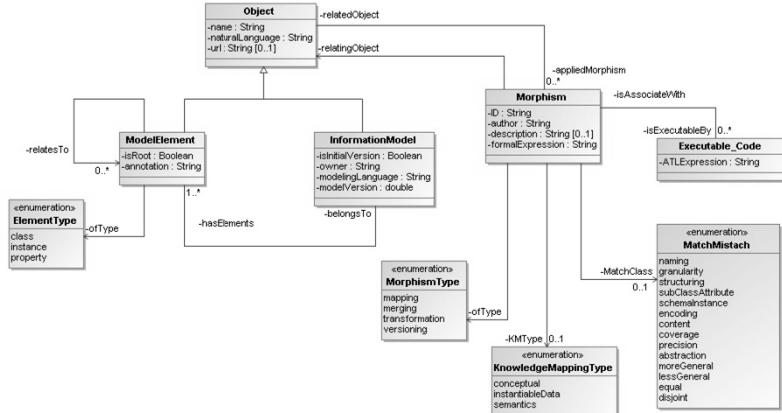


Fig. 4. Structure of knowledge base for communication support (CM)

The CM has two main classes: *Object* and *Morphism*. The *Object* represents any *InformationModel* (IM) which is the model/ontology itself and *ModelElements* (also belonging to the IM) that can either be classes, properties or instances. The *Morphism* basically represents the MapT described in the previous section: it associates a pair of *Objects* (related and relating – *Melems* in MapT), and classifies their relationship with a *MorphismType*, *KnowledgeMappingType* (if the morphism is a mapping), and *Match/Mismatch* class (*MatchClass* in MapT). The *Morphism* is also prepared to store transformation oriented *ExecutableCode* that will be written in the Atlas Transformation Language (ATL) and can be used by several organizations to automatically execute the mapping, transforming and exchanging data with their business partners as envisaged in [28].

5 Case Study from Mechanical Industry

The simple choice of a “bolt” supplier by a mechanical engineer/designer, very often brings interoperability issues. Suppliers usually define proprietary nomenclatures for their products and its associated knowledge representation (whatever format). Thus, the need to align product data and knowledge emerged as a priority to solve the dilemma. A possible solution is to allow each enterprise involved to keep its terminology and classification in use, and define formal mappings that mediate the communications among them using the tuple and CM here proposed (Figure 5).

The presented case study is related to a Client System able to represent and act as a retailer of “bolts” which has two different suppliers (Enterprises A and B). Thus, following the previously presented approach it was established a set of mappings between this Client’s information model and the others from the suppliers. Such mappings were defined following the MapT and were then recorded in the CM. In the left and right parts of the Figure 5 it is illustrated the Client, Enterprise A and Enterprise B

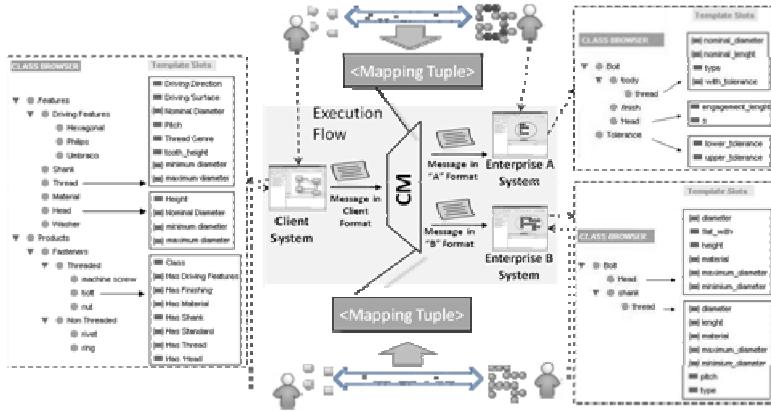


Fig. 5. Mapping design and execution flow in data exchange

information structures (in this case, ontologies). Using them it was possible to specify the set of tuples that define a traceable mapping at the information model level (level $n+1$), and therefore provide the support to establish/re-establish interoperability (i.e. sustain). Indeed, following the model-driven development paradigm, data transformations (level n) can be executed automatically and repeated whenever desired/demanded with little costs to the overall network interoperability (Figure 6) [28][29].

5.1 Case Study Tuples for Sustainable Interoperability

Taking specifically the example related to the tolerance characteristic of a bolt it is stated that in the client system such tolerance characteristic is defined by two properties, *maximum diameter* and *minimum diameter*. These are equally used by the Enterprise B. However, Enterprise A, uses the concepts *upper tolerance* and *lower tolerance*, which represents the same expected result but using different data values. *Nominal diameter* and *diameter* concepts have the same value and semantics in all the ontologies. Thus, the transformations equations related to the tolerance properties from Client to Enterprise A, are the following:

$$\text{maximum diameter} = \text{nominal diameter} + \text{upper tolerance} \quad (2)$$

$$\text{minimum diameter} = \text{nominal diameter} - \text{lower tolerance} \quad (3)$$

Due to space constraints, the authors decided to include only 3 tuples: one for each *KMType*. Taking the mapping example that relates the conceptual level of the *maximum diameter* and *upper tolerance* concepts respectively from Client and Enterprise A systems, the resulting tuple is specified as in Table 3 (upper section). As an example of MapT definition related to the structure with *KMType = Semantics*, it was chosen the relationship between a Client's bolt and Enterprise B's Bolt (Table 3 - middle section). They use different structures for the same concept, but the mapping

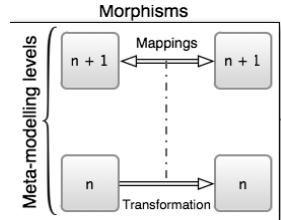


Fig. 6. Model Driven Framework (from [28])

is not absolute (e.g. B does not have driving features information). Finally, as an example of MapT $KMType = InstantiableData$, it was chosen the equation 2. In this case it is needed to define an operation using two data concepts from the related model: the *upper tolerance* and *nominal diameter*.

Table 3. Case Study MapT's

<i>ID</i>		<i>Client2.1.1.2.7.8_1</i>
<i>MEllems</i> = (a, b)	<i>a</i>	$((Thread)Products.Fasteners.Threaded.bolt.HasThread).maximum\ diameter$
	<i>b</i>	$((Tolerance)Bolt.body.thread.with_tolerance).upper_tolerance$
<i>KMType</i>		<i>Conceptual</i>
<i>MatchClass</i>		<i>MoreGeneral</i>
<i>Exp</i>		" $a \supseteq b$ "
<i>ID</i>		<i>Client2.1.1.2_1</i>
<i>MEllems</i> = (a, b)	<i>a</i>	<i>Products.Fasteners.Threaded.bolt</i>
	<i>b</i>	<i>Bolt</i>
<i>KMType</i>		<i>Semantics</i>
<i>MatchClass</i>		<i>Coverage</i>
<i>Exp</i>		" $a \sim b$ "
<i>ID</i>		<i>Client2.1.1.2.7.8_2</i>
<i>MEllems</i> = (a, b)	<i>a</i>	$((Thread)Products.Fasteners.Threaded.bolt.HasThread).maximum\ diameter$
	<i>b</i>	$Bolt.body.thread.nominal_diameter$
<i>KMType</i>		<i>InstantiableData</i>
<i>MatchClass</i>		<i>Granularity</i>
<i>Exp</i>		" $a = b + ((Tolerance)Bolt.body.thread.with_tolerance).upper_tolerance$ "

With the above mapping tuples specified, a simple example can be drawn to illustrate how the network could sustain its interoperability status: if for some reason Enterprise A started using English metric units (inches instead of millimetres), an *Encoding* mismatch would be detected and all mappings reused just by adapting the *InstantiableData* tuples according to the proper conversion rule. For example, the expression (*Exp*) of the last MapT would change to $a = (b + ((Tolerance)Bolt.body.thread.with_tolerance).upper_tolerance) * 25,4$, and the information in the CM updated automatically to enable a swift recovery of the interoperable status using the automatic code generation of model-driven capabilities. In the case of mappings not represented using the MapT format or in any other traceable form, enterprises would need to go through the full integration process again, reprogramming communication functions manually at huge costs.

6 Concluding Remarks and Future Work

The net effect of cheap communications is a perception that individuals and organizations have to deal in a world that is increasingly dynamical, complex and uncertain, and that their actions may have unintended consequences that impact on others [30]. Indeed, business networks are plagued with uncertainty, and systems interoperability

has become an important topic of research in the last years not only from the perspective of establishing it, but also of sustaining it.

The development of standards and ontologies helps to master the communication within those networks. However, alone, information standards do not solve today's enterprise interoperability problems. Indeed, typically each stakeholder has its own business requirements and suffers external influences that might lead to a harmonization breaking phenomena [9]. Therefore, organizations from similar business environments are still having trouble cooperating at a long term.

To address this issue and support sustainable interoperability in the context of collaboration networks, it is required to analyse how intervention strategies on the network evolution, namely attempts to shape local interaction patterns and mappings, affect the network interoperability sustainability. The authors use a knowledge enriched tuple for mappings representation that is stored on a communication mediator that keeps traceability of model mapping changes so that readjustments can be easier to manage, and data exchange re-established automatically using the model-driven development paradigm.

As for part of the future work is the validation of the CM for other types of morphisms, such as composition of models, merging operations, etc. For model altering morphisms, also the proposed tuple might need some readjustments, or even assume a different format. The major complexity associated to the study of the properties of complex systems is that the associated models, drive to non-linearity, which in turn, drives to difficulties in the system's study and in predicting their behaviour. In this context and also part of future work, at the system microscopic level, prediction could be seen as a proposal for the automatic adaptation of the network morphisms. Thus, the CM envisages to have associated learning capabilities, monitoring, diagnostic and prognostic services based on the operations history and interventions on the involved systems.

References

1. Friedman, T.: *The World is Flat*. Farrar, Straus & Giroux (2005)
2. Wilkinson, I., Young, L.: On cooperating: firms, relations and networks. *Journal of Business Research* 55(2), 123–132 (2002)
3. Amin, A., Cohendet, P.: *Architectures of knowledge: firms, capabilities, and communities*. Oxford University Press, Oxford (2004)
4. Camarinha-Matos, L., Afsarmanesh, H.: Collaborative networked organizations: a research agenda for emerging business models. Springer, Heidelberg (2004)
5. Jardim-Goncalves, R., Agostinho, C., Malo, P., Steiger-Garcia, A.: Harmonising technologies in conceptual models representation. *International Journal of Product Lifecycle Management* 2(2), 187–205 (2007)
6. Peppard, J., Rylander, A.: From Value Chain to Value Network: Insights for Mobile Operators. *European Management Journal* 24(2-3), 128–141 (2006)
7. Ray, S.R., Jones, A.T.: Manufacturing interoperability. *Journal of Intelligent Manufacturing* 17(6), 681–688 (2006)
8. White, W.J., O'Connor, A.C., Rowe, B.R.: Economic Impact of Inadequate Infrastructure for Supply Chain Integration. *NIST Planning Report 04-2*. Gaithersburg (2004)
9. Agostinho, C., Jardim-Goncalves, R.: Dynamic Business Networks: A Headache for Sustainable Systems Interoperability. In: Meersman, R., Herrero, P., Dillon, T. (eds.) *OTM 2009 Workshops. LNCS*, vol. 5872, pp. 194–204. Springer, Heidelberg (2009)

10. Axelrod, R.: *The Complexity of Cooperation: Agent-Based Models of Competition and Collaboration*. Princeton University Press, Princeton (1997) ISBN 0-6910-1567-8
11. Gharejadaghi, J.: *Systems Thinking: Managing Chaos and Complexity: A Platform for Designing Business Architecture*. Butterworth-Heinemann, Butterworths (2005) ISBN 0-7506-7973-5
12. Sugihara, G., May, R.M.: Nonlinear Forecasting as a Way of Distinguishing Chaos from Measurement Error in Time Series. *Nature* 344, 734–741 (1990)
13. Sarraipa, J., Jardim-Goncalves, R., Steiger-Garcao, A.: MENTOR: an enabler for interoperable intelligent systems. *International Journal of General Systems* 39(5), 557–573 (2010)
14. Lubell, J., Peak, R.S., Srinivasan, V., Waterbury, S.C.: Step, Xml, And Uml: Complementary Technologies. In: Proc. of DETC 2004, ASME 2004 Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Salt Lake City, Utah USA, September 28- October 2 (2004)
15. Guarino, N., Schneider, L.: Ontology-driven conceptual modeling. In: Pidduck, A.B., Mylopoulos, J., Woo, C.C., Ozsu, M.T. (eds.) CAiSE 2002. LNCS, vol. 2348, p. 3. Springer, Heidelberg (2002)
16. Agostinho, C., Delgado, M., Steiger-Garcao, A., Jardim-Goncalves, R.: Enabling Adoption of STEP Standards Through the Use of Popular Technologies. In: 13th ISPE International Conference on Concurrent Engineering (CE 2006), September 18-22 (2006)
17. INTEROP NoE. Deliverable MoMo.2 - TG MoMo Roadmap. InterOP (2005)
18. Agostinho, C., Sarraipa, J., D'Antonio, F., et al.: Enhancing STEP-based interoperability using model morphisms. In: 3rd International Conference on Interoperability for Enterprise Software and Applications, I-ESA 2007 (2007)
19. Delgado, M., Agostinho, C., Malo, P., Jardim-Gonçalves, R.: A framework for STEP-based harmonization of conceptual models. In: 3rd International IEEE Conference on Intelligent Systems (IEEE-IS 2006), Westminster, July 4-6 (2006)
20. INTEROP NoE. Deliverable D.A3.3 - Semantic Annotation language and tool for Information and Business Processes. InterOP (2006)
21. Delgado, M.: Harmonisation of STEP and MDA conceptual models using Model Morphisms. MSc thesis presented at Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa (2008)
22. Dauben, J.W., Cantor, G.: *His Mathematics and Philosophy of the Infinite*. Harvard University Press, Boston (1979)
23. Bernstein, P.A.: Applying Model Management to Classical Meta Data Problems. In: First Biennial Conference on Innovative Data Systems Research, California (January 2003)
24. Giunchiglia, F., Yatskevich, M., Shvaiko, P.: Semantic Matching: Algorithms and Implementation. *J. Data Semantics* 9, 1–38 (2007)
25. Blaha, M., Premerlani, W.: A catalog of object model transformations. In: Proceedings of WCRE, vol. 96, p. 87 (1996)
26. Nicolis, G., Prigogine, I.: *Exploring Complexity: An Introduction*. W. H. Freeman and Company, New York (1989)
27. Sarraipa, J., Zouggar, N., Chen, D., Jardim-Goncalves, R.: Annotation for Enterprise Information Management Traceability. In: Proceedings of IDETC/CIE ASME 2007 (2007)
28. Agostinho, C., Correia, F., Jardim-Goncalves, R.: 17th ISPE International Conference on Concurrent Engineering (CE 2010), Cracow, Poland, September 6-10 (2010) (accepted)
29. Selic, B.: The Pragmatics of Model-Driven Development. *IEEE Software Magazine* (September/October 2003)
30. Merali, Y., McKelvey, B.: Using Complexity Science to effect a paradigm shift in Information Systems for the 21st century. *Journal of Information Technology* 21(4), 211–215 (2006)

Planning and Scheduling for Dispersed and Collaborative Productive System

Marcosiris A. O. Pessoa, Fabrício Junqueira, Paulo E. Miyagi,
and Diolino J. Santos Fo

University of São Paulo, São Paulo, Brazil
{marcosiris,fabri,pemiyagi}@usp.br, diolino.santos@poli.usp.br}

Abstract. The advances of production technologies reflect a worldwide trend for sustainability and rational use of resources. As a consequence, some movements for industry reorganization can be observed in geographically dispersed systems with new productive systems (PSs) configurations. These PSs work with small and medium size production lots, product families with increasing variety, and hard delivery date. In this context, concepts of dispersed and collaborative productive systems (DCPSs), with new requirements for production scheduling are expected to be properly explored for improvement of PSs performance. Therefore, we here consider that productive activities can be treated as services and the SOA (service-oriented architecture) approach can be adopted for the integration of the DCPS. Then, a planning service based on time windows and a scheduling service that uses the APS (advanced planning and scheduling) heuristics is proposed to assure the expected performance of the DCPS.

Keywords: disperse productive system, planning and scheduling, scheduling heuristics, time windows.

1 Introduction

The international competitive pressure on manufacturing enterprises has strongly increased. Nowadays, consumers and delivery date orient products demands and the lots of products have small and medium size, while product families are increasing in variety [1]. Hence, technological advances and market changes have established new efficiency patterns in the production of products [2]. Moreover, there is the sustainability scope; customers are demanding sustainable green products and processes [3]. The use of the Internet can collaborate with the sustainability reducing personal displacement, travel costs, and the related carbon footprints [3]. In this context works that contributes for practical implementation of dispersed and collaborative productive systems (DCPS) are also a contribution for sustainability.

The DCPS are collections of autonomous machines connected and integrated through a communication network. The focus is on a method for integrating equipment and machine control strategies to ensure the accomplishment of productive processes [2], [4]. However, in these works, issues related from the planning and scheduling viewpoint are not treated; in others words, the optimization of where and when services should be available is not addressed. In this context, this work introduces a “planning service” based on “time windows heuristics” and a “scheduling service” to assure the expected performance of the PS.

Several works adopted service-oriented architecture (SOA), in which Web Service (WS) is a popular instance of this architecture [2], [4], [5], and [6]. In [2] and [4] was introduced a method to specify the productive processes aiming at the automation and coordination of their activities and services. In this sense, this work adopts the concept of SOA for DCPS, and the “planning service” allow customers to know if their orders are feasible at the requested delivery date without the need of a complete review of the scheduling. The “scheduling service” allocates the services assuring that they will be delivered at the requested time. The Petri net is used for modeling the integration of services and to evaluate the proposal.

This paper is organized as follows. Section 2 presents its contribution to sustainability. Section 3 presents the key concepts used in its development. Section 4 presents the DCPS with the planning and the scheduling service and a model in which these services are integrated with other parts of the PS. Finally, section 5 presents the conclusions.

2 Contribution to Technological Innovation for Sustainability

According to [7], living with “sustainability” requires a strong engagement of science, industry and politics. “Sustainability research” is expected to lead with three conflicting aspects: (i) contributing to economic development, (ii) being ecologically acceptable, and (iii) being socially fair. The reference [3] states that sustainable green engineering design and manufacturing are changing every aspect of our life and have already originated a wide area of research topics. One of these topics is about machine and process optimization, i.e., using real-time monitoring systems, with sustainability statistics integrated into the feedback-controlled system; it is possible to avoid out-of-control situations in any process step. Another topic is the Internet communication, which can be used in DCPS to reduce human displacement, direct and indirect production costs and the related carbon footprint. In accordance with some works ([8], [9], and [10]), the teleoperation approach, for example, causes social impacts such as: personnel trip reduction for meetings, operator activities improvement in unfavorable conditions, or dangerous environments, among other applications.

Besides the topics above mentioned, this work also contribute to waste reduction. In fact, considering a production environment, i.e., a DCPS in which each PS produces items for other PSs, the problem of planning and scheduling has no trivial solution and cannot be neglected. The delay in delivery of any intermediary item may directly interfere with the delivery date of the final product and the rational use of resources can be impaired.

The adopted SOA approach for DCPS that includes “planning service” and “scheduling services” is fundamental to improve the performance of the overall system, and represent a contribution to sustainability.

3 Fundamental Concepts

According to [2], the evolution of the Internet velocity and functionalities reduced the implementation cost of DCPSs and also increased teleoperated systems flexibility. These systems have computers called “clients” with applications that communicate

with other applications located at other computers named “servers”, which control remote equipment or resources directly. With this “client/server” structure, multiple operators can use, via Internet, equipment or resources that are geographically distributed in other places. Operators and equipment are installed in geographically dispersed locations and interact to accomplish the desired tasks [11].

This work uses the “time windows” [12] approach and the “constraint programming mechanism” [13] to plan the process; these two approaches are encapsulated at “planning services”. For modeling the integration between services, the Petri net is used. Below is a description of “time windows” and “constraint programming mechanism”.

3.1 Time Windows

In production problems with delivery priority dates, it is possible to determine, for each production lot, the latest instant of time at which this lot should be ready. This latter instant becomes the maximum delivery date for a production lot. The “time windows” are time intervals for the allocation of production lots; in other words, in a time scale, they are the limits at which the lots can be allocated (due dates), and each lot has its own “time windows”. The four key parameters of a “time windows” (Fig.1) are:

- Earliest start time (*Est*) is the earliest time at which a lot can be started in the “time windows”;
- Latest finish time (*Lft*) is the latest time at which a lot must be finalized in the “time windows”;
- Earliest finish time (*Eft*), is the earliest time at which a lot can be finalized in the “time windows”, and it is defined as the sum of *Est* with the process time (*Pt*) of the lot;
- Latest start time (*Lst*), is the latest time at which a lot must be started in the “time windows”, and it is defined as the *Lft* subtracted from the *Pt* of the lot.

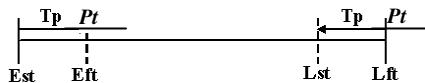


Fig. 1. The four key parameters of a “time windows”

The “time windows” limits must be consistent with the restrictions of (*Est* and *Lft*) for all lots, and for this reason, a “constraint programming mechanism” is used. In this work, the following restrictions are considered:

- Capacity constraints [12] and [13] that ensure that the production lots do not overlap in time, i.e., it is not possible to process two lots simultaneously at the same resource;
- Precedence constraints by product recipe that ensures that: the *Est* of the consumer lot must be equal to or greater than the *Eft* of the producer lot, and the *Lft* of the producer lot must be smaller than or equal to the *Lst* of the consumer lot;
- Storage restrictions [14], related to the storage capacity.

3.2 Modeling Services Using Petri Net

The Petri net is a powerful modeling tool, and several works have used it to model dispersed systems ([2], [4], [15], [16], [17] and [18]). Thus, the reader is assumed to be familiar with the basic concepts of Petri net [19]. In [15], a survey into process modeling is presented. Verification techniques (based on Petri net, process algebra and state machine) are described as well as the tools developed to ensure the specification and composition of services via Internet. Relevant issues concerning Petri net and process algebra are also presented in [16], in which the graphical representation of Petri net is highlighted. In [17], the authors compare two semantics of Petri nets for the orchestration of services via Internet. The Petri net is used herein for modeling and evaluating the integration of services in DCPS.

4 DCPS with Planning and Scheduling Services

This study adopts the approach based on SOA (service oriented architecture) presented in [20] for coordinating the implementation of DCPS. Customers, operators and PSs are considered to be geographically dispersed. The adopted structure is multi-layered with three levels (Fig.2). At the top level (presentation layer), there is the service that exposes the functionality of DCPS for a communication network (Internet) and provides mechanisms for planning and scheduling the customers orders considering the available resources. At this level, the “time windows” and the “constraint programming mechanism” are used in the “planning services”, and the “earliest” APS (advanced planning and scheduling) heuristic proposed in [21] is used for scheduling the services in the “scheduling services”. In the intermediate layer (integration and coordination layer) there are the mechanisms for integration and coordination of services for PSs involved in the production process. The lower layer (productive services layer) treats the PS services that may be required to meet the orders. The possibility to reconfigure the service structures of each PS is considered.

In the DCPS proposed, the customers have access to the “customers interface”, so that they can send requests to a repository of products through communication devices (PDAs, netbooks, web sites, 3G phones, etc.) and also monitor the state of the requested services. The interface (“management request services”) accounts for managing the orders and for informing customers about the order feasibility. This service has interfaces with the “customers interface”, the data base, and the “planning services”. If the order is feasible, the “planning services” sends the considered “time windows” to the “scheduling services” and also sends a message to the “management request services” confirming that the order is feasible. Using APS heuristics, the “scheduling services” allocates the order and sends a message to the “integration and coordination services”. The architecture depicted in Fig.2 considers that global processes are centrally coordinated by an orchestrator for the integration and coordination of services. This orchestrator is realized by the “integration and coordination services” that coordinates the activities of all PSs involved in the execution of the processes. Each PS of the DCPS is encapsulated in one of the service “teleoperative system of production services”, which exposes functionalities of the manufacturing system module as a service enabling the interaction between the module and the “integration and coordination services”.

Fig.3 presents the Petri net model of the presentation layer, the integration and coordination layer and their relationship. The graph describes the flow of messages and the services that are activated to execute the production orders.

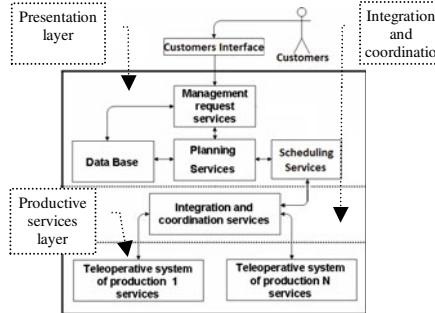


Fig. 2. DCPS architecture with “planning services” and “scheduling services”

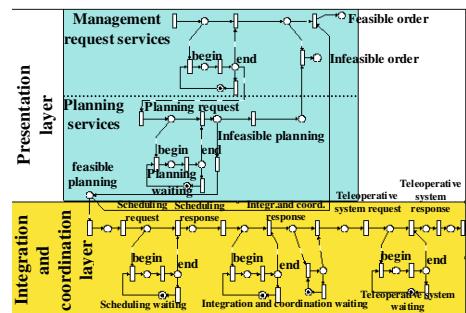
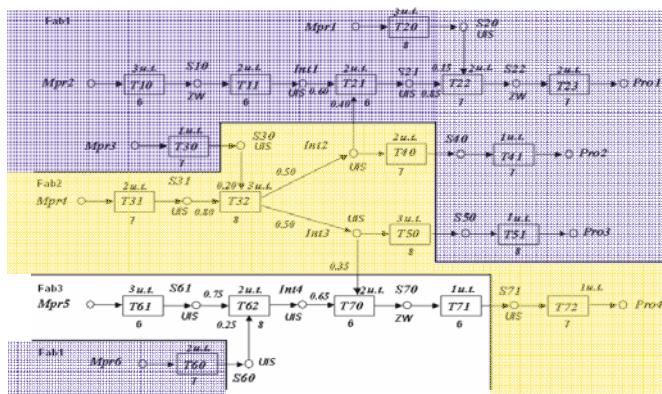


Fig. 3. Petri net Model of the integration of the “presentation layer” and the “integration and coordinating layer”

This model was developed with the support of Petri net tools (software packages such as HPSim [22]) that was also used for structural and behavior analysis based on the properties of the graph.

5 Results

The example (Fig. 4) considers the processing of the bill of material ([23]) to be produced in DCPS that operates in lots. It involves different lot sizes, successive operations, storage conditions and limited shared resources. This section uses the example presented in [18]. This is composed of a set of three PSs of the same DCSP, *Fab1*, *Fab2* and *Fab3*. There is a scheduling sector that remotely manages the scheduling of



all PSs. *Fab1* provides products *Pro1*, *Pro2* and *Pro3*, and *Fab2* provides *Pro4*. Input *S30* is produced by *Fab1* and consumed by *Fab2*. Fig. 4 shows the STN (state-task-network) [24] of the PSs process.

The PSs are geographically dispersed, but the STN was used to demonstrate the relationship between processes and that the delayed delivery of *S30* may cause delay in delivery of products *Pro1*, *Pro2*, *Pro3* and *Pro4*. In Tab.1 and Tab.2 are presented respectively, the assignment of tasks to the resources, the demand for final products and delivery dates. The storage restrictions are only for items *S10*, *S22* and *S70*.

Table 1. Tasks Assignment

Productive Systems (PS)	Resources	Tasks
<i>Fab1</i>	<i>P1</i>	<i>T10, T21, T51</i>
	<i>P4</i>	<i>T23, T30, T60</i>
	<i>P7</i>	<i>T11, T22, T41</i>
<i>Fab2</i>	<i>P2</i>	<i>T32</i>
	<i>P3</i>	<i>T31, T72</i>
	<i>P5</i>	<i>T20, T40, T50</i>
<i>Fab3</i>	<i>P6</i>	<i>T61, T70</i>
	<i>P8</i>	<i>T62, T71</i>

Table 2. Demand for final products and due date

Productive Systems (PS)	Products	Clients	Quantity	Delivery Date
<i>Fab1</i>	<i>Pro1</i>	<i>Client1</i>	70	64
	<i>Pro2</i>	<i>Client2</i>	50	88
	<i>Pro3</i>	<i>Client3</i>	50	86
	<i>S30</i>	<i>Fab2</i>	10	35
<i>Fab2</i>	<i>S60</i>	<i>Fab3</i>	20	14
	<i>Pro4</i>	<i>Client4</i>	50	67
	<i>Int2</i>	<i>Fab1</i>	32	44
	<i>Int3</i>	<i>Fab3</i>	32	44
	<i>S20</i>	<i>Fab1</i>	16	26
	<i>S40</i>	<i>Fab1</i>	56	66
	<i>S50</i>	<i>Fab1</i>	56	69
	<i>S71</i>	<i>Fab2</i>	54	55
<i>Fab3</i>				

Fig.5 presents the “time windows” generated by the “planning services” and the scheduling generated by the “scheduling services”. The horizontal line represents the time of the scheduling horizon.

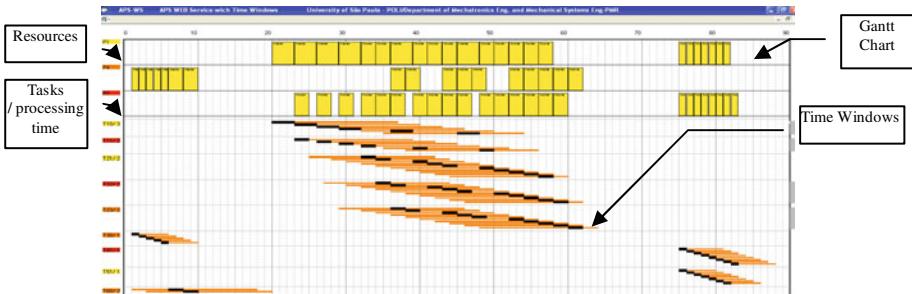


Fig. 5. The “time windows” and scheduling of *Fab1* orders

In this example was used the APS (advanced planning and scheduling) heuristic presented in [21].

6 Conclusions

An architecture based on services towards a modular and scalable design of dispersed and collaborative productive system (DCPS) is presented. In addition, the use of a heuristic based on “time windows” and the “constraint programming mechanism” for

“planning services” was proposed. Thus, “planning services” was developed and implemented as a web service, based on “time windows” that avoid improper task assignments. This service allows customers or others factories (productive systems) to know if their request is feasible without the need of a complete review of the scheduling. Thus the customer has a faster response about the feasibility of the order request. The “scheduling services” uses the earliest APS heuristic, which assures that the order will be delivered at the desirable time. The DCPS here considered assures the performance of the activities related to manufacturing products even if the production process is fragmented into several productive subsystems. The performance of the DCPS was analyzed by using the Petri net model and study cases.

Acknowledgments. The authors would like to thank the Brazilian governmental agencies CAPES, CNPq and FAPESP, for their partial support to this work.

References

1. Yamaba, H., Matsumoto, S., Tomita, S.: An Attempt to Obtain Scheduling Rules of Network-Based Support System for Decentralized Scheduling of Distributed Production Systems. In: 6th IEEE Intern. Conf. on Industrial Informatics (INDIN), Daejeon, Korea, pp. 506–511 (2008)
2. Garcia Melo, J.I., Junqueira, F., Morales, R.A.G., Miyagi, P.E.: A Procedure for Modeling and Analysis of Service-Oriented and Distributed Productive Systems. In: 4th IEEE Conf. on Automation Science and Engineering (CASE), Washington, DC, USA, pp. 941–946 (2008)
3. Ranky, P.G.: An Integrated Architecture, Methods and Some Tools for Creating More Sustainable and Greener Enterprises. In: IEEE International Symposium on Sustainable Systems and Technology (ISSST), Arlington, VA, USA, pp. 1–6 (2010)
4. Garcia Melo, J.I., Junqueira, F., Miyagi, P.E.: Towards Modular and Coordinated Manufacturing Systems Oriented to Services. DYNA 77(163), 201–210 (2010)
5. Mönch, L., Zimmermann, J.: Providing Production Planning and Control Functionality by Web Services: State of the Art and Experiences with Prototypes. In: 5th IEEE Conf. on Automation Science and Engineering (CASE), Bangalore, India, pp. 495–500 (2009)
6. Spiess, P., Karnouskos, S., Guinard, D., Savio, D., Baecker, O., Souza, L.M.S., Trifa, V.: SOA-Based Integration of the Internet of Things in Enterprise Services. In: IEEE International Conference on Web Services (ICWS), Los Angeles, CA, USA, pp. 968–975 (2009)
7. Zickler, A., Mennicken, L.: Science for Sustainability: the Potential for German-Brazilian Cooperation on Sustainability-Oriented Research and Innovation. In: 1st German-Brazilian Conf. on Research for Sustainability, São Paulo, SP, Brazil (2009)
8. Lee, J.S., Zhou, M., Hsu, P.L.: A Petri-Net Approach to Modular Supervision with Conflict Resolution for Semiconductor Manufacturing Systems. IEEE Trans. on Automation Science and Engineering 4(4), 584–588 (2007)
9. Backes, P.G., Tso, K.S., Norris, J.S., Steinke, R.: Group Collaboration for Mars Rover Mission Operations. In: IEEE Intern. Conf. on Robotics & Automation (ICRA), vol. 3(1), pp. 3148–3154 (2002)
10. Marble, J.L., Bruemmer, D.J., Few, D.A.: Lessons Learned from Usability Tests with a Collaborative Cognitive Workspace for Human-Robot Teams. In: IEEE Intern. Conf. on System, Man, and Cybernetics (SMC), vol. 1(1), pp. 448–453 (2003)
11. Eilhajj, I., Xi, N., Fung, W.K., Liu, Y.H., Hasegawa, Y., Fukuda, T.: Modeling and Control of Internet Based Cooperative Teleoperation. In: IEEE Intern. Conf. on Robotics and Automation (ICRA), vol. 1(1), pp. 662–667 (2001)

12. Sadeh, N.: Look-Ahead Techniques for Micro-Opportunistic Job Shop Scheduling. PhD Thesis, School of Computer Science, Carnegie Mellon University, Pittsburg, Pennsylvania, USA (1991)
13. Caseau, Y., Laburthe, F.: Improving Branch and Bound for Job-Shop Scheduling with Constraint Propagation. In: 8th Franco-Japanese Conf., Brest, France (1995)
14. Rodrigues, L.C.A., Magatão, L., Setti, J.A.P., Laus, L.P.: Scheduling of Flow Shop Plants with Storage Restrictions Using Constraint Programming. In: 3rd Intern. Conf. on Production Research, Americas' Region (ICPR-AM), Brazil (2006)
15. Breugel Van, F., Koshkina, M.: Models and Verification of BPEL. Technical Report (2006),
<http://www.cse.yorku.ca/~franck/research/drafts/tutorial.pdf>
16. Van der Aalst, W.M.P.: Pi Calculus Versus Petri Nets: Let Us Eat ‘humble pie’ Rather than Further Inflate the “pi hype”. BPTrends 3(5), 1–11 (2005)
17. Lohmann, N., Verbeek, E., Ouyang, C., Stahl, C.: Comparing and Evaluating Petri Net Semantics for BPEL. Computer Science Report, Technische Universiteit Eindhoven (2007)
18. Pessoa, M.A.O., Garcia Melo, J.I., Junqueira, F., Miyagi, P.E., Santos Fo, D.J.: Scheduling Heuristic Based on Time Windows for Service-Oriented Architecture. In: 8th Intern. Conf. on Mathematical Problems in Engineering, Aerospace and Sciences (ICNPAA), Sao Jose dos Campos, SP, Brazil (2010)
19. Hruzand, B., Zhou, M.C.: Modeling and Control of Discrete Event Dynamic Systems. Springer, London (2007)
20. Garcia Melo, J.I., Fattori, C.C., Junqueira, F., Miyagi, P.: Framework for Collaborative Manufacturing Systems Based in Services. In: 20th Intern. Congr. of Mechanical Engineering (COBEM), Gramado, RS, Brazil (2009)
21. Rodrigues, M.T.M., Gimeno, L., Pessoa, M.A.O., Montesco, R.E.: Scheduling Heuristics Based on Tasks Time Windows for APS systems. Computer Aided Chemical Engineering 18, 979–984 (2004)
22. Kin, S.Y.: Modelling and Analysis of a Web-Based Collaborative Information System - Petri Net-Based Collaborative Enterprise. Intern. Journal of Information and Decision Sciences 1(3), 238–264 (2009)
23. Systems Instrumentation and Automation Society. Enterprise-Control System Integration: Part1: Models and Terminology. 95.00.01 (2000)
24. Papageorgiou, L.G., Pantelides, C.C.: Optimal Campaign Planning/Scheduling of Multi-purpose Batch/Semicontinuous Plants - part 2 - A Mathematical Decomposition Approach. Industrial and Engineering Chemistry Research, 35(2), 510–529 (1996)

An SOA Based Approach to Improve Business Processes Flexibility in PLM

Safa Hachani¹, Lilia Gzara¹, and Hervé Verjus²

¹ G-SCOP Laboratory, Grenoble Institute of Technology, France

² LISTIC Laboratory, University of Savoie, France

{safa.Hachani,lilia.Gzara}@grenoble-inp.fr

herve.verjus@univ-savoie.fr

Abstract. Companies collaborating to develop new products need to implement an effective management of their design processes (DPs). Unfortunately, PLM systems dedicated to support design activities are not efficient as it might be expected. DPs are changing, emergent and non deterministic whereas PLM systems based on workflow technology don't support process flexibility. So, needs in terms of flexibility are necessary to facilitate the coupling with the environment reality. Furthermore, service oriented approaches (SOA) ensure a certain flexibility and adaptability of composed solutions. Systems based on SOA have the ability to inherently being evolvable. This paper proposes an SOA based approach to deal with DP flexibility in PLM systems. To achieve this flexibility, the proposed approach contains three stages. In this paper we focus on the first stage "identification".

Keywords: PLM system; Business processes; flexibility; SOA.

1 Introduction

To stay competitive, companies are adopting IT solutions to facilitate collaborations and improve their product development. Among these IT solutions, Product Lifecycle Management (PLM) systems play an essential role by managing product data. Furthermore, one of the goals of PLM is to foster collaboration among different actors involved in product development processes [1]. Thus, each PLM system provides (1) a database to store product information and the functions necessary to the management of this stored data and (2) integrate a tool to model, execute and control business processes (BPs) associated to product design. Most PLM systems are adopting workflow management solutions to cope with BPs. Nevertheless, brakes analysis of design processes (DPs) support in PLM systems points out a lack of flexibility for change; justified by the stiffness of formal workflow models and workflow systems used in PLM [2]. However, in a context where the organizations are in a constant seek of balance facing up a highly volatile environments; work methods (BPs) cannot be fixed definitively, especially, when dealing with DPs which are emergent. Furthermore, various hazards intervene during DPs due to external constraints (such as customer requirements evolution, supplier constraints, etc.) and/or internal constraints (such as technical feasibility problems, staff absence, etc.) Thus, DPs are characterized by their

instability, relative inconsistencies of the rules that govern them and their incompleteness [3]. As a result, companies face several obstacles, including the limited implementation of new work methods. So, needs in terms of flexibility are necessary to facilitate the coupling with the business reality. Following these two findings: i) instability of DPs models and ii) rigidity of workflow technology, we identify a critical need to deal with DPs flexibility.

Furthermore, software engineering evolved towards new paradigms such as services oriented approaches (SOA) based on services composition. Composition approaches ensure a certain flexibility and adaptability of composed solutions [4]. Systems based on SOA have the ability to inherently being evolvable [5]. Therefore, some standards development organizations have been involved in the development of standards for PLM with SOA [6, 7]. Much attention has been directed towards the use of SOA with PLM [8, 9, 10]. However, the main objective of current studies on SOA in PLM systems is to enable the online integration of heterogeneous PLM systems in order to enhance partner's collaboration (when many companies collaborate in product design and industrialisation). Otherwise, in PLM domain, there is no work that focused on BP agility using SOA. Although much work has been done to date, more studies need to be conducted to deal with BP agility using SOA.

To deal with BPs flexibility in PLM, we propose an approach that makes profiles from SOA. The objective is to specify, design and implement DPs in a flexible way so they can rapidly adapt to changing conditions. In this paper, we present the approach and their three stages and then we focus on the first stage; "service identification".

The remainder of this paper is organized as follows. Section 2 presents our contribution to Sustainability. Section 3 illustrates how dynamic process change happens in PLM systems with a motivation example. Section 4 present existing approaches dealing with PLM and SOA. Section 5 presents the SOA based approach for business process flexibility. Section 6 presents the functional PLM services identification approach. Conclusion remarks and future work are given in Section 7.

2 Contribution to Sustainability

Changing BPs can be considered both at design time and at runtime. Evolvable BPs stands for processes that may be adapted on-the-fly (i.e. while they are enacted and executed). So, we only focus on approaches being able to apply changes without the need to redefine the whole process model. Recent investigations deals with service orientation that promotes light-coupling, services reuse and dynamic composition that cope with sustainability. Loosely coupled services may be organized and composed according to needs and expectations. Dynamic services composition stands for assembling and re-assembling services while the process is executing. So, by proposing evolvable business process based on services technologies, the change can be done without need to redefine whole process model. Moreover, changes will be always done on the unique version of process model with reusing deployed services. It means the system doesn't need to archive unused version. Furthermore, as change is done by reusing deployed services, we may reduce the consumed energy of developing new functionalities and process models. Services expose functionalities as operations independently of their real implementation and can be reused even if the implementation is changing (as consequence, some changes do not affect the services composition).

3 Motivation Example

We propose to use an example of process managed in PLM systems to illustrate the implementation and dynamicity issues of DPs on these systems. This example describes the process of assigning production orders to production workshops according to illustration of figure 1 (a). It includes several activities: visualizing of commands' catalog, creating production order, consulting workshops timetable, assigning production order. In order to be managed on PLM system, this process should be automated. The first step consists on creating the whole functionalities necessary to the fulfillment of the process and storing them on the PLM database. Ones the functionalities stored, workflow template should be defined. Finally, the workflow can be instantiated from the template and executed till its completion [11].

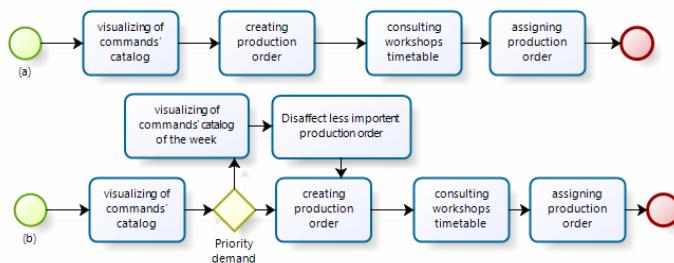


Fig. 1. Change in the assigning production orders process

As this BP is executed several times, the process model may change due to practical situations. In fact, the actual change is driven by many factors, such as technology shift, internal changes or external changes, etc., and they cannot be anticipated at design time. Some of these factors, for example, external changes (i.e. in order to satisfy suppliers or partners needs), may only cause temporary changes of a business process. While, some factors, such as regulation change, may cause permanent changes. Nevertheless, the BPs in PLM systems are in permanent change due to the dynamic environment justified by the diversity of suppliers, project's partners and requirement changes (for instance, time and cost constraints). For example, some production orders may be urgent; the customer requires a time constraint. The workflow process will evolve to the one shown in Figure 1 (b). Ones process changes occur, new workflows templates are defined and workflows instances are initiated accordingly. Defining new workflows templates requires creating and storing the new functionalities in the PLM database. Thus, the deployment of new functionalities can take much time. In addition it's necessary to handle the previous workflows instances which are initiated from old workflow templates. Most workflow systems, used in PLM consider two steps; before applying the new workflow template, the old instance has to be stopped and restarted according to the new workflow template. Indeed, after restarting the workflow instance some completed tasks have to be carried out unnecessarily (for example, creating production order). To address the above problem, dynamic business process management in PLM system might be brought in as a potential solution. We need a method which facilitates change on business process by (1) reducing the time lost on the deployment of necessary functionality and (2) eliminating repetitive execution of completed tasks.

4 Related Works

Currently there are two initiatives in terms of PLM services, OMG PLM services and OASIS PLCS web services. We studied these proposed standards in order to decline a SOA vision applied in PLM domain and to list existing PLM services. PLM Services is an OMG standard specification [6]. It was developed to implement, operate and support online access and batch operation using several international standards. It's based on PDTnet standard [12] which defines mechanisms to query and traverse instances of the data schema defined for the PLM services specification. This standard proposes a set of services to launch the execution of PDTnet queries which provide the necessary computational functionalities to create, read, update, and delete instances of data. The second standard, OASIS PLCS PLM Web Services [7] is an ISO STEP standard. PLCS is an information exchange standard that provides a number of functional modules. Each one targets a specific area of PLM information. These modules provide services such as searching for PLM business objects or loading information objects. Each one addresses a set of business objects (nouns) and a number of services (verbs) that operate on them. The standard verbs are: Create, Remove and Update. A first analysis of these two standards enables us to conclude that their main focus is related to the problem of online integration of heterogeneous PLM systems to implement their collaborative processes. Furthermore, this study has highlighted a first track for services we target to use to achieve workflow's activities.

Few researches have used these two standards to handle PLM system issues. Erkan Gunpinar [8] have used OMG PLM service standard to interface two PLM systems. His objective is to use PLM Services standard for PLM data exchange via internet in order to speed up collaborative business process done between two heterogeneous PLM systems. With the same objective, Dag Bergsjö [9] proposed a framework to support ECM along with two developed KBE applications that simulate effects of a change in real time, as the product is updated in the PLM system. To do that, he separates the uppermost layers (applications used in the business process to create and process information) and the bottom layers (legacy systems which store information, such as PLM systems) with a connector based on PLM services. Kim [10] used OMG PLM services to introduce MEMPHIS a data exchange middleware that provide common interfaces and enable a centralized integration of multiple PLM systems (server) from any point (clients). All above researches focused on the integration of heterogeneous PLM systems to allow collaboration and did not address the problem of dynamic change of business processes in PLM. So, dynamic business process change remains an unsolved problem in PLM system.

5 Adopted Method

BPs flexibility can be perceived differently. The meaning it takes determines the way to handle the flexibility topic. From our part, BPs flexibility is the fast reactivity to internal and external changes and the easiness to modify BPs models and to set up the new business activity. This perception of process flexibility arises the need to get a method which allows composing evolvable BP models. We propose to enrich the expression of BP models and open the way for modeling by dynamic service

orchestration. This supposes that once change happens, we can add to, delete from or replace an activity with another. The challenge here is to address the mechanism needed for a solid implementation of dynamic BP change on a real PLM system.

To fulfil these expectations, we resort to service orientation [4, 5, 13]. SOA is a promising solution answering these expectations since it allows building agile and interoperable enterprise systems. So, if services represent a good answer to technical level issue, they can be used to handle the business level issues: service can be directly invoked by business users and executed as basic steps of BPs. Services can be combined and reused quickly to meet business needs. SOA promotes light coupling between services which can be dynamically combined in order to support agility. The services are defined as providers of reusable business functions in an implementation independent function that is loosely coupled to other business functions [12]. Indeed, SOA organizes the basic functionality contained in the systems on a set of services, which can be combined and reused to meet business needs. This vision therefore allows the construction of new systems by reusing existing services; which called services composition [14]. Thus, the concept of service as previously described as a loosely coupled piece of functionality can be composed and reused to quickly respond to BP change and to achieve the new model without needing to replace it completely and to re-execute completed tasks. So, we resort to service oriented approach in order to propose reusable activities as business services and evolvable business processes as business services composition. A business PLM service exposes a business activity needed to support a business need of a product DPs. It specifies corresponding features (functional PLM service) necessities to the development of this activity. Afterward, we dynamically compose identified business PLM services in order to implement on a flexible way the articulations of business. Moreover, to insure the alignment between technical level and business level they should be a mechanism that allows execution of identified business PLM services with the same language chosen for the business level. Thus, we propose a set of functional PLM services that represent the whole possible features of PLM system. So, once a new functionality is needed to perform change, operations of functional PLM services can be solicited from the database (service repository) to do it.

In order to achieve this solution, we have to complete three stages. Propose an approach for service identification: steps, techniques and criteria necessary to the identification of business and functional services catalog. Propose services based modeling paradigm for dynamic BP definition. Propose alignment technique that allows moving from business level to technical level. This technique ensures a continuum of transformation from specification to implementation of BPs. In this paper we concentrate only on the first stage; service identification stage.

6 Service Identification Stage

Our objective is to offer two catalogs of services; business PLM service catalog expressing the business needs of product DP and functional PLM services catalog enabling the execution of business needs. To define the services catalogs, we defined the appropriate techniques and necessary steps to achieve service identification stage. Thus, we conducted an initial pass of SOA development approaches.

6.1 Service Identification Approach

Several approaches are interested on the development of SOA [15-18]. All proposed approaches are based on Service Oriented Modeling Architecture (SOMA) [15]. SOMA is an analysis and design method used for the design and construction of SOA. It focuses on techniques for the identification, specification, and realization of services. All studied work present what activities should be carried out to develop an SOA and how each activity should be conducted. Especially, we focused on service identification activity and studied the proposed service identification techniques. We have identified two complementary approaches used to identify candidate services: Top-Down approach and Bottom-up approach. A Top-Down approach offers a mapping of business use cases, which means to separate the business domain into major functional areas and subsystems. Then functional areas are decomposed on sub-processes, and high-level services. This technique is called domain decomposition. While a Bottom-Up approach analyzes the existing systems to identify low-level services. This approach is called Existing asset analysis.

So, we propose an hybrid approach to identify the two catalogs. We propose to deal with a top-down approach to define business services needed to achieve business process and a bottom-up approach to identify functional PLM services. In this paper we concentrate on the functional PLM service identification method.

6.2 Functional PLM Services Identification Method and Catalog

A functional PLM service is a collection of PLM operations which reflects functions expected by PLM system users. Each operation implements the concept of automated business task and exposes a function of PLM. We propose a bottom-up approach based on three steps for Functional PLM Service identification: (i) Identifying PLM data categories, (ii) Identifying operations of each category and (iii) Grouping identified operations on functional PLM services. Therefore, some criteria are needed to help decide which operations can be grouped together; functional dependencies and process dependencies. Below we detail this approach throw Functional PLM service identification.

To identify the functional PLM services operations, two kinds of information sources were used following the proposed identification approach. On the one hand, two PLM systems were examined to identify the product data categories and their related operations. Thus we have identified a first truck of operations offered on PLM system. For instance, Display product structure, Compare BOMs, etc. On the other hand, we have organized meeting with business expert to validate and enrich the list of identified operations. The results of this two first step are shown on Fig. 2.

The last step of the identification approach is to classify the identified operations on functional PLM services. This classification is done by testing functional dependencies and process dependencies criteria between the identified operations by using dependencies matrix in which column headers are the same as the corresponding row header and they correspond to the identified operations. Functional dependencies are those between operations have a common global purpose. In this case each matrix entry (a_{ij}) correspond to the couple (OperationPurpose i, OperationPurpose j). While,

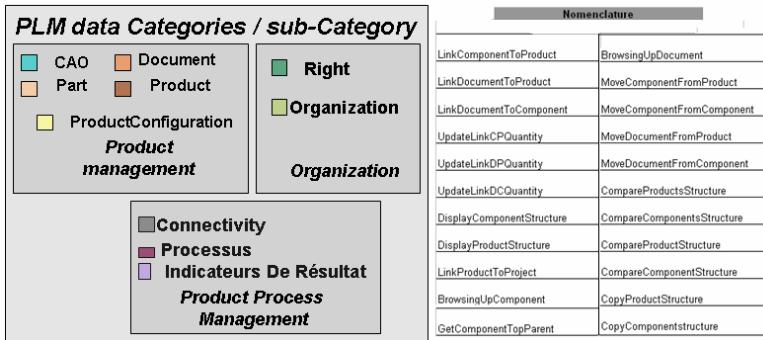


Fig. 2. On the left: Cartography of PLM data categories. On the right: An excerpt from Functional PLM Services operations list.

processing dependencies are those between services that are choreographed together to make up a high level service; used together frequently. In this case each matrix entry (a_{ij}) corresponds to the couple (OperationHighLevelGoal i, OperationHigh-LevelGoal j). Thus, if an entry has the same couple's element it means that there is dependency between it corresponding row and column operations. To group identified operations, the final decision is done by the superposition of two dependency matrix. After all we have grouped Validate parts, Build Combinations and Develop BOM operations, etc. on a same functional PLM service named Manage Product Configuration. Moreover, we have grouped ReviewDatapack, DistributeReviewDatapack, and NotifyReviewRequestor operations on another functional PLM service named Management of Design Review.

7 Conclusion

In this paper we discussed the problem of BPs flexibility on PLM system. To allow dynamic BP change in PLM system, an approach based on service technology is introduced. This paper proposes to deal with BP model as a business PLM services composition. The challenge here is to react quickly to changes either by replacing some services by other ones or by adding new services to the composition. In order to deal with alignment issues between technical and business level, we propose a service type for each level (functional and business). Business PLM services reflect the business needs of different stakeholders and will be executed by a composition of functional PLM services which meet the functionalities of PLM system. In this paper we concentrated on the identification stage for functional PLM Service. At this stage, a reflexive posture is going to be conducted to define the business service concept, its identification method in order to propose the business service catalog. Moreover, we will address techniques that allow moving from one level to another level to ensure a continuum of transformation from specification to implementation of BPs.

References

1. Sääksvuori, A., Immonen, A.: Product lifecycle management. Birkhäuser, Basel (2005)
2. Bowers, J., Button, G., Sharrock, W.: Workflow from within and without. In: Proceedings of the Fourth European Conference on CSCW, pp. 51–66. Kluwer, Dordrecht (1995)
3. Green, P., Rosemann, M.: Integrated process modelling: an ontological evaluation, information systems. *Information Systems*, 73–87 (2000)
4. Papazoglou, M.P.: Service-oriented computing: Concepts, characteristics and directions. In: CS, I. (ed.) WISE 2003, pp. 3–12 (2003)
5. Kontogiannis, K., Lewis, G.A., Smith, D.B.: The landscape of service-oriented systems: A research perspective. In: Proc.s of Int. Workshop on Systems Development in SOA Environments (2006)
6. Lämmer, L., Bugow, R.: PLM Services in Practice. *The Future of Product Development*, 503–512
7. Srinivasan, V.: An integration framework for product lifecycle management. In: Computer-Aided Design (in Press, Corrected Proof)
8. Gunpinar, E., Han, S.: Interfacing heterogeneous PDM systems using the PLM Services. *Advanced Engineering Informatics* 22, 307–316 (2008)
9. Bergsjo, D., Catic, A., Malmqvist, J.: Implementing a service-oriented PLM architecture focusing on support for engineering change management. *International Journal of PLM* 3, 335–355 (2008)
10. Kim, S.R., Weissmann, D.: Middleware-based Integration of Multiple CAD and PDM Systems into Virtual Reality Environment. *Computer-Aided Design & Applications* 3, 547–556 (2006)
11. Khoshafian, S., Buckiewicz, M.: Groupware et workflow. InterÉditions (1998)
12. Credle, R., Bader, M., Brikler, K., Harris, M., Holt, M., Hayakuna, Y.: SOA Approach to entreprise integration for product lifecycle management, pp. 66–80 (October 2008)
13. Erl, T.: Service-Oriented Architecture (SOA): Concepts, Technology, and Design. Prentice Hall, Englewood Cliffs (2005)
14. Manouvrier, B., Ménard, L.: Intégration applicative EAI, B2B, BPM et SOA. Hermès Science (2007)
15. Arsanjani, A., Allam, A.: Service oriented modeling and architecture for Realization of an SOA. In: On The Proceeding of IEEE Int. Conf. on Service Computing (SCC 2006). IEEE, Los Alamitos (2006)
16. Papazoglou, M.P., Heuvel, W.-J.: Service-oriented design and development methodology. *Int. J. of Web Engineering and Technology (IJWET)* 2, 412–442 (2006)
17. Zimmermann, O., Kroghdahl, P., Gee, C.: Elements of Service-Oriented Analysis and Design (2004), <http://www.ibm.com/developerworks/library/ws-soad1/>
18. Chang, S.H., Kim, S.D.: A Service-Oriented Analysis and Design Approach to Developing Adaptable Services. In: The IEEE International Conference on Services Computing (SCC 2007), pp. 204–211 (2007)

A Model for Automated Service Composition System in SOA Environment

Paweł Stelmach, Adam Grzech, and Krzysztof Juszczyszyn

Wrocław University of Technology, Institute of Computer Science, Wybrzeże
Wyspiańskiego 27, 50-370 Wrocław, Poland

{Pawel.Stelmach, Adam.Grzech, Krzysztof.Juszczyszyn}@pwr.wroc.pl

Abstract. In this paper a holistic approach to automated service composition in Service Oriented Architecture is presented. The model described allows to specify both functional and non-functional requirements for the complex service. In order to do so a decomposition of the complex service composition process into three stages is proposed, in which a structure, a scenario and an execution plan of a complex service is built. It is believed that service composition tool based on the proposed model will be able to create complex services satisfying efficiently both functional and nonfunctional requirements.

Keywords: service oriented architecture, service composition, SOA.

1 Introduction

The current trend in business is increasingly leading to the use of external services. The number of such services is growing and will grow even faster in the future. Soon delegated services will not be limited to simple tasks performed by individual providers, but they will be complex services performing complex business processes.

Business process designers are rarely experts in web-services technology. More often they can specify the business process in the form of functional requirements that are abstract to the underlying software infrastructure. As a consequence a service composition system should focus on formalization and interpretation of the functional and non-functional requirements in order to find services that fulfill them at a given time and are not hardcoded into the business logic.

The composition of Web services makes that a reality by building complex workflows and applications on the top of the SOA model [5, 8, 3, 13]. However, literature shows a wide range of composition approaches [1, 2]. In [4] it is mentioned that, rather than starting with a complete business process definition, the composition system could start with a basic set of requirements and in the first step build the whole process, whereas many approaches (i.e. [6]) require a well-defined business process to compose a complex service.

Current work often raises the topic of semantic analysis of user requirements, service discovery (meeting the functional requirements) and the selection of specific services against non-functional requirements (i.e. execution time, cost, security). Review of the literature indicates, however, that these methods have not yet been

successfully combined to jointly and comprehensively solve the problem of composition of complex services that satisfy both functional and non-functional requirements. In many cases only one aspect is considered. For example [9] focuses on services selection based only on one functional requirement at a time. [6, 14, 15, 16, 17] show that non-functional requirements are considered to be of a key importance, however many approaches ignore the aspect of building a proper structure of a complex service which is key to optimization of i.e. execution time. Many AI Planning-based approaches ([7]) focus on functionalities of the complex service but leave no place for the required non-functionalities satisfaction.

The following sections will present an approach to service composition which models functional and non-functional requirements (in section 3) and in section 3 the decomposition of the service composition task that fulfills both functional and non-functional requirements. Section 5 shows an example to visualize the composition process.

2 Contribution to Sustainability

Service Oriented Architecture (SOA) is an approach which prolongs the life of applications, increasing their reusability by publishing them as web services which can be accessed independent of the system (i.e. its programming language) that wants to use them. They are accessible through their web-enabled interfaces by the means of XML message system. In the case of SOA legacy applications or their parts can be searched and executed whenever functionalities are needed but those functionalities have to be properly modeled in order to conduct a successful search. In this context automatic service composition allows for sustainability of the system through the re-use of various web-services in a form of complex web-services.

3 Complex Service Composition Problem Definition

3.1 Service Level Agreement

The Service Level Agreement (SLA) is a contract that specifies user requirements:

$$SLA_l = (SLA_{lf}, SLA_{lnf}) \quad (1)$$

where SLA_{lf} are the functional requirements, SLA_{lnf} are the non-functional requirements and l stands for a l -th complex service request.

The functional requirements have to be fulfilled by functionalities offered by a set of atomic services. Those services composed in a specific structure form a complex service fulfilling both functional and non-functional requirements (like availability, performance etc.).

Functional requirements consist of request of functionalities and some time dependencies among them:

$$SLA_{lf} = (\Phi_l, R_l) = (\{\varphi_s, \varphi_{l1}, \varphi_{l2}, \dots, \varphi_{lk}, \dots, \varphi_k\}, R_l) \quad (2)$$

where:

- Φ_l is a set of functionalities φ_{li} , in which φ_s and φ_k are starting and ending functionalities
- R_l is a set of time dependencies among functionalities such that:

$$r_{ijl} = \begin{cases} 1 & \text{when } \varphi_{li} \prec \varphi_{lj} \\ 0 & \text{otherwise} \end{cases} \text{ and } r_{ijl} \in R_l, \text{ where } i, j \in \{1, 2, \dots, n_l\}$$
- n_l is a number of required functionalities in SLA_{lf}

Non-functional requirements are represented by a set of constraints:

$$SLA_{lnf} = \{\psi_{l1}, \psi_{l2}, \dots, \psi_{lk}, \dots, \psi_{lm_l}\} \quad (3)$$

where ψ_{li} is a requirement of a single non-functionality – a constraint put on the composed complex service (i.e. required execution time of the whole service) and m_l is a number of required non-functionalities in SLA_{lnf} .

3.2 Service Composition Task

The service composition task is formulated as follows:

for a given

- request of a complex service SLA_l
- repository of services AS: as_k is a k -th atomic service,
- family of execution graphs $\{G(V, E)\}$ representing all possible execution plans of the requested complex service in which vertices contain services ($as_k \in V$) and edges from E define the succession of services in a complex service,

find an optimal execution plan $G(V, E)$ of a complex service, fulfilling the functional and non-functional requirements by minimizing the quality criterion:

$$G(\{as_1^*, as_2^*, \dots, as_k^*\}, E^*) \leftarrow \min_{as_1, as_2, \dots, as_k, E} Q(G(\{as_1, as_2, \dots, as_k\}, E), SLA_l) \quad (4)$$

where:

- $as_1^*, as_2^*, \dots, as_k^*$ are the selected optimal services that belong to corresponding k -th vertex of graph G (each vertex contains one service that fulfills one functional requirement), where optimal means that all services fulfill functional requirements and no other set of services gives better non-functional properties of the complex service (given a set of edges E)
- E^* is an optimal set of edges that determines the succession of services $as_1^*, as_2^*, \dots, as_k^*$. Optimal means here that no other set of edges gives better non-functional properties of the complex service and that the set of edges is minimal, meaning no edge could be rejected without the loss of consistency of the graph.

4 Problem Decomposition

Service composition process can be described as a series of graph transformations beginning with the transformation of users requirements (SLA_l) to the graph form and

ending with an execution graph representing the complex service fulfilling the functional and non-functional requirements of the SLA_l. The service composition problem can be decomposed into three stages:

- **structure:** SLA_l is given a graph form, where all functionalities are embedded in vertices and time dependencies are expressed in the form of edges (preferably in the minimal number of edges possible without loss of information),
- **scenario:** the set of edges of the graph is extended so that the graph is consistent (complete information about order of execution of all functionalities is known) and service repository is searched in order to find services capable of providing requested functionalities,
- **execution plan:** for each vertex only one service is chosen (from candidates gathered in the previous step), so that all the services in the structure fulfill the non-functional requirements.

4.1 Complex Service Structure Stage

At this stage users requirements are transformed into the form of the graph. Each vertex contains one functionality required by the user and each edge represents the order dependency between two requirements (services that fulfill them). In the resulting graph there can be less edges than would appear from the set of time dependencies R_l . This is because a simpler structure is preferred if the simplification does not change the correct order defined in R_l . Fig. 1 depicts the situation where in the left side there is a graph with dependencies defined by the user while graph on the right after simplification has a serial structure (with edge (c) omitted). All the dependencies are still valid.



Fig. 1. Reduction of complexity of the structure

The task is formulated as follows:

for a given $SLA_{lf} = (\{\varphi_s, \varphi_{l1}, \varphi_{l2}, \dots, \varphi_{lk}, \dots, \varphi_k\}, R_l)$
find:

$$GB_l(VB_l, EB_l = EB_l^*): EB_l^* = \arg \min_{EB_{lk}} \|EB_{lk}\| \quad (5)$$

where:

- $VB_l = \{v_s, vb_{l1}, vb_{l2}, \dots, vb_{lk}, \dots, vb_{ln_l}, v_k\}$, $v_s = \varphi_s$, $vb_{li} = \varphi_{li}$, $v_k = \varphi_k$
- $EB_l \ni eb_{ijl}$, such that $\forall_{r_{ijl} \in R_l}$ if $r_{ijl} = 1$, then exists a path $d(vb_i, vb_j)$

4.2 Complex Service Scenario Stage

Adding missing edges. The graph obtained at the previous stage rarely is consistent. However, there cannot be any functionalities that have no connection with the rest of

the graph, because this would result in them not being executed at all. The task is formulated as follows:

for a given $GB_l(VB_l, EB_l)$ **find** a consistent graph $GC_l(VC_l, EC_l = EC^*)$ such that:

$$EC_l^* = \arg \min_{EC_{ls}} d_{max}(GC_l(VC_l, EC_{ls}), vb_s, vb_k) \quad (6)$$

where:

- $d_{max}(GC_l(VC_l, EC_{ls}), vb_s, vb_k)$ – length of the longest path from vertex vb_s to vertex vb_k in graph $GC_l(VC_l, EC_{ls})$
- $VC_l = VB_l$
- $EB_l \subset EC_{ls} = EB_l + EA_{ls}$

The criterion in (6) ensures that structures with a higher degree of parallelization are preferred. This will directly influence the non-functional properties of the complex service i.e. execution time.

Services selection. Based on the users functional requirements atomic services are selected from the services repository. Services are found after semantic and structural comparison of their descriptions and functionalities from graph GC_l . The problem of services selection could not approached in depth in this paper. For more information please refer to [10] and [11]. The task of services selection is formulated as follows:

for a given

- consistent graph $GC_l(VC_l, EC_l)$ (in each vertex containing functional requirements)
- service repository AS

find:

- graph $GS_l(VS_l = \{AS_{l1}, AS_{l2}, \dots, AS_{lk}, \dots, AS_{ln_l}\}, ES_l)$ of valid realizations (services fulfilling the functional requirements) such that:

$$\exists_{j \in \{1, 2, \dots, J\}} as_j = as_{lk_u} \in AS_{lk} \text{ if } \varphi(as_{lk_u}) \supset \varphi_{lk} \text{ for each } u = \{1, 2, \dots, u_{lk}\} \quad (7)$$

where:

- as_{lk_u} is a service that fulfills requirement φ_{lk} and u_{lk} is a **number** of services that fulfill that requirement (and will replace it in appropriate vertex AS_{lk} in graph GS_l)
- $\varphi(as_{lk_u})$ are functionalities offered by the service and $\varphi(as_{lk_u}) \supset \varphi_{lk}$ means that service as_{lk_u} offers more or equal functionalities than required
- $AS_{lk} = \{as_{lk_1}, as_{lk_2}, \dots, as_{lk_{u_{lk}}}\}, AS_{lk} \subset AS, AS_{lk} \in VS_l$

4.3 Complex Service Execution Graph

After the scenario construction stage, each vertex of graph GS contains a number of services that satisfy the corresponding functional requirements. To obtain the execution graph G_l in each node of graph GS_l one service has to be selected, such that – with regard to the edges that create the structure of the graph – the complex service consisting of selected services fulfills the non-functional requirements. The task is formulated as follows:

for given

- graph $GS_l(VS_l = \{AS_{l1}, AS_{l2}, \dots, AS_{lk}, \dots, AS_{ln_l}\}, ES_l)$,
- non-functional requirements $SLA_{lnf} = \{\psi_{l1}, \psi_{l2}, \dots, \psi_{lk}, \dots, \psi_{lm_l}\}$,
- weight w_{lk} of k-th non-functionality of SLA_{lnf} ,

find

- an optimal execution graph $G_l(V_l, E_l)$ such that:

$$G_l(V_l, E_l) \leftarrow \min_{G_{lp}(V_{lp}, E_l)} \sum_{k=1}^{m_l} w_{lk} [\psi_{lk} - f_{\psi_k}(G_{lp}(V_{lp}, E_l))] \quad (8)$$

where:

- $G_{lp}(V_{lp}, E_l) \in FG_l(V_l, E_l)$ – family of valid execution graphs (each vertex contains one service selected from $AS_{lk} : V_{lp} \ni v_{lpk} = \{as_j \in AS_{lk}\}$ and the set of edges stays the same as in GS_l graph: $E_l = ES_l$)
- m_l – length of vector of non-functional requirements Ψ_l of l-th service request
- $f_{\psi_k}(G_{lp}(V_{lp}, E_l))$ is a function aggregating non-functionalities (corresponding to ψ_k) of atomic services embedded in the execution graph $G_{lp}(V_{lp}, E_l)$

with subject to the following constraints:

$$\forall_{k=1}^m: f_{\psi_k}(G_{lp}(V_{lp}, E_l)) \leq \psi_{lk} \quad (9)$$

5 Example

For given:

$$SLA_{lf} = (\{\varphi_1, \varphi_2, \varphi_3\}, R_l), \text{ where } r_{ijl} \in R_l: r_{ijl} = \begin{cases} 1 & \text{for } i = 1 \text{ and } j = 2 \\ 0 & \text{otherwise} \end{cases}$$

(for simplicity φ_s and φ_k are omitted in this example)

$$SLA_{lnf} = \Psi_l = [7, 4]; w_{l1} = 1, w_{l2} = 2, \psi_{l1} - \text{cost}, \psi_{l2} - \text{time},$$

Service repository $AS = \{as_j\}, j = 1, 2, \dots, J$, where:

$$\begin{aligned} \varphi(as_1) &= \varphi_1, \varphi(as_2) = \varphi_2, \varphi(as_3) = \varphi_2, \varphi(as_4) = \varphi_3, \varphi(as_5) = \varphi_3, \dots \\ \psi(as_1) &= [2, 1], \psi(as_2) = [3, 2], \psi(as_3) = [2, 2], \psi(as_4) = [1, 4], \psi(as_5) = [3, 1] \end{aligned}$$

Find optimal complex service execution graph $G_l(V_l, E_l)$ that fulfills the SLA.

Solution:

In the **structure** stage from the given SLA_{lf} only one possible can be obtained:

$$GB_l = GB_l(\{\varphi_1, \varphi_2, \varphi_3\}, \{(\varphi_1, \varphi_2)\})$$

In the **scenario** stage there are numerous possible consistent graphs (Fig. 3). As (6) ensures the more parallel structure is preferred (here: $d_{max}(GC_l^*, vb_s, vb_k) = 1$) and the optimal graph GC_l^* is obtained:

$$GC_l^* = GC_{ln}(\{\varphi_1, \varphi_2, \varphi_3\}, \{(\varphi_1, \varphi_2), (\varphi_1, \varphi_3)\})$$

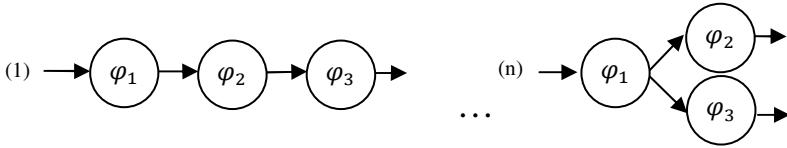


Fig. 2. Possible scenarios of a complex service

In step two of scenario stage for each functional requirement services that fulfill it are selected from the repository AS and so graph GS_l is as follows:

$$GS_l(\{AS_1 = \{as_1\}, AS_2 = \{as_2, as_3\}, AS_3 = \{as_4, as_5\}\}, \{(AS_1, AS_2), (AS_1, AS_3)\})$$

In the **execution graph** stage based on the non-functional parameters of selected service an optimal execution graph can be found:

$$G_l(V_l^*, E_l) = G_l(v_{l1}^* = \{as_1\}, v_{l2}^* = \{as_3\}, v_{l3}^* = \{as_5\}, \{(v_{l1}^*, v_{l2}^*), (v_{l1}^*, v_{l3}^*)\})$$

For the above execution graph G the value of quality criterion (8) is minimal:

$$\sum_{k=1}^2 w_{lk} [\psi_{lk} - f_{\psi_k}(G(V^*, E_l))] = 1 * (7 - 7) + 2(4 - 3) = 2$$

6 Conclusions

The model for service composition presented in this paper allows for specification of both functional and non-functional requirements. The decomposition of the task allows for specification of three problems: transformation of user requirements into the structure of the graph, creating a scenario for the complex service by making the graph consistent and selecting services and at last finding the execution graph for the service that fulfills all requirements. The proposed approach takes into consideration the fact that constructing an appropriate structure for the complex service and selection of services with various non-functionalities are crucial to the problem of fulfilling non-functional requirements and should not be solved separately.

During the next stage of development the security assessment of atomic services will be taken into account, according to formal approach recently presented in [12]. This will allow to address security as an important non-functional property of complex services, barely covered in recent research and practical applications.

A composition framework based on the presented model is developed and will be made available in the form of package of tools for service composition.

The incoming stage of the project also assumes the application and evaluation of the framework in an industrial scenario.

Acknowledgments

The research presented in this paper has been partially supported by the European Union within the European Regional Development Fund program no. POIG.01.03.01-00-008/08.

References

1. Jinghai, R., Xiaomeng, S.: A Survey of Automated Web Service Composition Methods. In: Cardoso, J., Sheth, A.P. (eds.) SWSWPC 2004. LNCS, vol. 3387, pp. 43–54. Springer, Heidelberg (2005)
2. Milanovic, N., Malek, M.: Current Solutions for Web Service Composition. *IEEE Internet Computing* 8(6), 51–59 (2004)
3. Charif, Y., Sabouret, N.: An Overview of Semantic Web Services Composition Approaches. *Electronic Notes in Theoretical Computer Science*, 146, 33–41 (2006)
4. Aggarwal, R., Verma, K., Miller, J., Milnor, W.: Constraint Driven Web Service Composition in METEOR-S. In: Proceedings of the 2004 IEEE International Conference on Services Computing, pp. 23–30 (2004)
5. Blanco, E., Cardinale, Y., Vidal, M., Graterol, J.: Techniques to Produce Optimal Web Service Compositions. *IEEE Congress on Services*, 553–558 (2008)
6. Ko, J.M., Kim, C.O., Kwon, I.-H.: Quality-of-service oriented web service composition algorithm and planning architecture. *The Journal of Systems and Software* 81, 2079–2090 (2008)
7. McIlraith, S., Son, T.: Adapting Golog for composition of semantic web services. In: Proceedings of 8th Intl. Conf. of Knowledge Representation and Reasoning (2002)
8. Ponnekanti, S.R., Fox, A.: SWORD: A developer toolkit for Web service composition. In: Proceedings of the 11th World Wide Web Conference, Honolulu, HI, USA (2002)
9. Klusch, M., Fries, B., Sycara, K.: OWLS-MX: A hybrid Semantic Web service matchmaker for OWL-S services. In: Web Semantics: Science, Services and Agents on the World Wide Web, vol. 7, pp. 121–133 (2009)
10. Stelmach, P., Prusiewicz, A.: An improved method for services selection. In: XVII International Conference on Systems Science, Wrocław (2010)
11. Stelmach, P., Juszczyszyn, K., Grzelak, T.: The scalable architecture for the composition of soku services. In: 31th International Conference on Information Systems, Architecture, and Technology, Szklarska Poręba (2010)
12. Kolaczek, G., Juszczyszyn, K.: Smart Security Assessment of Composed Web Services. *Cybernetics and Systems* 41(1), 46–61 (2010)
13. Agarwal, V., Chafle, G., Dasgupta, K., Karnik, N., Kumar, A., Mittal, S., Srivastava, B.: Synthy: A system for end to end composition of web services. In: Web Semantics: Science, Services and Agents on the World Wide Web. World Wide Web Conference 2005, Semantic Web Track, vol. 3(4), pp. 311–339 (2005)
14. Zeng, L., Kalagnanam, J.: Quality Driven Web Services Composition. In: 12th International Conference on the World Wide Web, pp. 411–421 (2003)
15. Huang, A.F.M., Lan, C., Yang, S.J.H.: An Optimal QoS-based Web Service Selection Scheme
16. Karakoc, E., Senkul, P.: Composing semantic Web services under constraints. *Expert Systems with Applications* 36(8), 11021–11029 (2009)
17. Ko, J.M., Kim, C.O., Kwon, I.-H.: Quality-of-service oriented web service composition algorithm and planning architecture. *Journal of Systems and Software* 81(11), 2079–2090 (2008)

The Proposal of Service Oriented Data Mining System for Solving Real-Life Classification and Regression Problems

Agnieszka Prusiewicz and Maciej Zięba

Institute of Informatics, Faculty of Computer Science and Management,
Wroclaw University of Technology, Wybrzeże
Wyspiańskiego 27, 50-370 Wrocław, Poland
`{Agnieszka.Prusiewicz,Maciej.Zięba}@pwr.wroc.pl`

Abstract. In this work we propose an innovative approach to data mining problem. We propose very flexible data mining system based on service-oriented architecture. Developing applications according to SOA paradigm emerges from the rapid development of the new technology direct known as sustainability science. Each of data mining functionalities is delivered by the execution of the proper Web service. The Web services, described by input and output parameters and the semantic description of its functionalities, are accessible for all applications that are integrated via Enterprise Service Bus.

Keywords: Service Oriented Data Mining, Sustainable design, SOA, Classification Services

1 Introduction

With the rising necessity of mining huge data volumes and knowledge discovery from many distributed resources there is a natural interest of using grid solutions. Execution machine learning algorithms in distributed environments allow organizations to execute computationally expensive algorithms on large databases, with relatively inexpensive hardware. Additionally an opportunity to merge data and discovered knowledge from many geographically distributed resources are created by the Internet. These elements favour of approving of a new field named as Distributed Data Mining (DDM) [11]. The survey of some Grid-based data mining systems is given in [1]. The other type of distributed computing, that rapidly develops in last decade is based on Service Oriented Architecture (SOA) paradigm [8]. The main idea of SOA is to treat applications, systems and processes as encapsulated components, which are called services. These services are represented by input and output parameters and the semantic description of its functionalities. Combining distributed data mining techniques with Web services has a lot of advantages. Web services are currently seen as a solution for integration of the heterogeneous resources and making heterogeneous systems interoperable. They are self-contained, self-describing and modular applications that can be published and invoked across the Web [15]. As an example of some Web services based data mining applications we can indicate Web based system for

metadata learning (WebDiscC) [13] or Association Rule Mining software called DisDaMin [2]. But the area of Web services-based data mining systems is still not well recognised. Taking into account current tendencies toward computer systems development there is a need for elaboration data mining distributed systems compatible with SOA paradigm.

In this work we propose Service Oriented Data Mining System (SODM System) for solving chosen data mining tasks i.e. classification tasks, equipped with mechanisms for advising which classifier should be used as the best for a given user request (data and requirements that must be classified). The advantage of our solution that is a consequence of compatibility with SOA paradigm, is that the users may use the model of classifiers and classifiers that have been created by the others. It is caused by implementing the functionalities of building models of classifiers and classifiers as Web services that are published and accessible via Internet.

2 Contribution to Sustainability

Presented in this paper Data Mining System is innovative approach to well-known machine learning solutions. Developing applications according to SOA paradigm emerges from the rapid development of the new technology direct known as sustainability science. Representing machine learning solutions as Data Mining Services has a significant contribution to sustainable design of new technological solutions. Data mining solutions are successively used for supporting decision-making processes in nature-society systems. They can be applied for extracting long-term trends in environments or to predict survival period in modern society. One of the drawbacks of data mining approaches is the problem with integration between such solutions and life-support systems from different fields of science. For instance, if there is a need to classify credit holders in bank system additional functionalities for customer modelling must be implemented in such system. SODM System solves the problem of integration and accessibility. In data mining solutions are fully accessible for all applications, which are integrated via ESB. For example, each application has an ability to invoke services responsible for creating clusters of objects simply by sending the objects in SOAP message. There is no need to create additional components of the application responsible for solving data mining problems (that is time consuming and demands data mining abilities), because these solutions are available as Data Mining Services in SODM System. Life-support systems, which communicate, with Data Mining Services are becoming interdisciplinary systems. Additionally, different systems representing various disciplines can invoke the same Data Mining Services. For instance, the same model can be used to predict survival period by medical and governmental system.

Every new solution developed by economists, biologists or mathematicians can be easily added to the SODM System. According to the basic sustainable design principle i.e. durability new data mining functionalities may be add as the new services without a need of rebuilding the overall system. This solution can be easily evaluated by data mining researchers, or simply used by all applications, which has access to ESB.

3 Service Oriented Architecture and Web Services

The basic concept of Service Oriented Architecture (SOA) approach is a semantically described service. In this context, SOA is the application framework that enables organizations to build, deploy and integrate these services independent of the technology systems on which they run. In SOA, applications and infrastructure can be managed as a set of reusable assets and services. The main idea about this architecture was that businesses that use SOA could respond faster to market opportunities and get more value from their existing technology assets [12]. According to SOA paradigm services are published and then access via Enterprise Service Bus (ESB) and used by Web applications. ESB is an implementation technology supporting SOA approach. ESB as an enterprise-wide extendable middleware infrastructure provides virtualization and management of service interactions, including support for the communication, mediation, transformation, and integration technologies required by services [10]. Web services are technologies that are based on XML (Extensible Markup Language) for messaging, service description and discovery. [8]. Web services use such standards as: SOAP, WSDL, and UDDI. SOAP (Simple Object Access Protocol) is a protocol for application communication and exchanging information over HTTP. WSDL (Web Services Description Language) is an XML-based language for describing Web services and how to access them. And finally UDDI (Universal Description, Discovery and Integration) is open standard for services discovery and invoking. UDDI being interrogated by SOAP messages provides access to WSDL documents that describe the protocol bindings and message formats required to interact with the Web services. In other words specifications define a way to publish and discover information about Web services. The relations between services provider and requester using xml standards are as follows. The provider gives an access to a Web service by publishing a WSDL description of its Web service, and the requester accesses the description using a UDDI or other type of registry that contains register of discovered services, and requests the execution of the provider's service by sending a SOAP message [8](Fig.1).

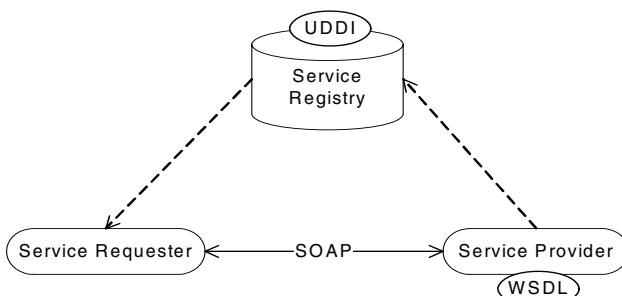


Fig. 1. The schema of interaction between Web services components

4 The Architecture of SODM System

The problems of classification are one of the common issues of data mining field [5,6,7,17]. There are plenty of dissertations, which touch mainly following problems: developing ensemble models for classification accuracy improvement, dealing with missing values of attributes, or building algorithms for incremental learning [3,4,14]. In most cases authors concentrate mostly on creating complex, difficult to understand methods, which can be used by data mining experts. Considering a classification task connected for example with labelling of web clients for marketing purposes there is a need for building a suitable model of classifier using past observations of the clients and their behaviours. The built model is further used to classify the new clients. Such model should be easily updated for a new set of data. To solve the classification tasks the application must be equipped with the mechanisms of building, updating and finding the most suitable model of the classifier. We propose SODM System that is based on Service Oriented Architecture (SOA) and uses encapsulation of data mining solutions in services. Hence each of the models of classifiers is represented by a service. In each service it is possible to distinguish operations responsible for creating and testing the model or classifying object using created model. The communication between applications (service clients) and service providers is made by Enterprise Service Bus (ESB) using Simple Object Access Protocol (SOAP). SODM System consists of the following components: User Interfaces, Service Usage Repository, Data Mining Services and Data Mining Services Manager (Fig. 2). All the components are integrated via ESB. User Interface is a component responsible for communication between users and services. It can be default application, which has ability to communicate with other Data Mining Components. More than one User Interface can be integrated with ESB. For instance using one interface it is possible to invoke Data Mining Services responsible for building and updating the model of classifier and the other can be used on mobile phone to classify unclassified object using this model. Service Usage Repository is responsible monitoring the usage of Data Mining Services. Data Mining Services are services responsible for solving classification and regression problems, clustering objects, data filtering or extracting associative rules. To improve the quality of usage of those services Data Service Manager is considered in the system. This service is responsible for filtering the most suitable services for the problem stated by the user in a request.

Data mining solutions presented in above architecture have couple of advantages. First of all, Data Mining Services are easily accessible by every service client, which is integrated via ESB. The client can be a mobile application, educational system, or some process in the system.

To invoke a Data Mining Service it is sufficient to create proper SOAP message, send it using ESB and interpret the response message. There is no need to implement data mining solutions locally because all such mechanisms are implemented and covered in the service. Moreover, the new data mining functionalities can be easily included in the system as a new service with specified input and output parameters. Presented architecture includes also Data Mining Services Manager, which helps to find the best solutions for users without data mining abilities.

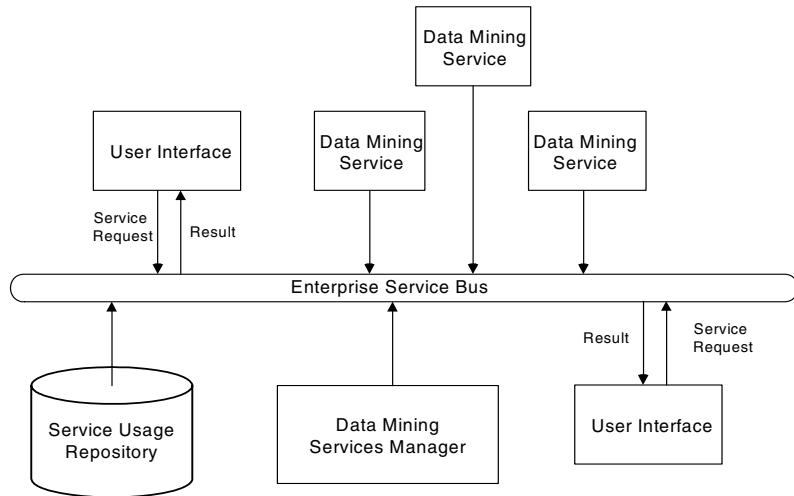


Fig. 2. The Components of SODM System

4.1 The Functionality of SODM System

Following functionalities can be distinguished in the SODM: building and updating classification and regression models, grouping objects using various clustering algorithms, filtering data (missing values of attributes replacement, features selection, etc.) and associative rules extraction. In this paper we concentrate only on functionalities related with solving classification and regression problems. In this field we can distinguish: building the model of classifier (training the classifier), instances classification, testing performance of the model of classifier, printing the model of classifier and updating the model. The problem of building the classifier is one of key issues in pattern recognition field. The model of classifier takes on the input vector of features values of the object to be classified and returns the class label (continuous value if regression problem is considered). This model can be given by an expert (as a set of rules) or it can be extracted using data composed of past observation of the objects (training dataset). In SODM System each of classifiers types (decision trees, decision rules, neural networks) is represented by one Data Mining Service (Classification Service) and building the classifier functionality is fulfilled by an operation which takes training dataset and returns the label of built model. Other functionalities are realized by other operations included in Classification Service. Instances classification is made using operation, which takes set of unlabeled objects on the input and returns their labels on the output by making classification using existing model. The performance of the model of classifier can be evaluated using testing operation. The dataset is given on the input of the service and testing methodology is defined and the operation returns testing rates values (accuracy of classification, Kappa statistic, etc.). Some of models of classifiers (rules, decision trees) have got understandable structure so there is an operation in each of Classification Service, which returns the structure of the model. There are couples of models of classifiers (Naïve Bayes), which can be

easily updated, in incremental learning process. If the model of classifier is updatable the corresponding Classification Service contains an operation responsible for updating the model.

SOSM System has additional functionalities related with filtering Classification Services. These functionalities are fulfilled by operations included in Classification Services Manager. Classification Service Manager is able to choose the models of classifier which are accurate for the given on the input dataset by fining corresponding Classification Services. There is also possibility for defining additional requirements for the model like updatability, which cannot be extracted from the data.

4.2 The Classification Components of SODM System

Classification Services

Following Classification Services are distinguished in our system: NBService, J48Service, JRipService, MLPService and LRSERVICE. These services are respectively represents following types of classifiers: Naïve Bayes, J48 (decision tree), JRip (decision rules), MLP (neural network) and LRSERVICE (Logistic Regression) [17]. Each of Classification Services contains operations described in previous subsection: buildClassifier (building the model of classifier), classifyInstances (classifying objects), getCapabilities (getting capabilities of the classifier), printClassifier (printing the structure of classifier), testingClassifier (testing the performance of classifier). Additionally, NBService contains the operation, which enables to update the model of classifier [14]. Initially SODM System contains only five Classification Services, but additional models of classifiers can be easily included in the system as service with defined standard operations and parameters.

Classification Services Manager

SODM System includes also managing service responsible for finding the best models according to request given by the client. For instance, the client can invoke the service by sending only the set of past observations, which should be used to build the model of classifier, and the response message will provide the most suitable Classification Services. Classification Service Manager contains three filtering operations. In first operation only dataset is taken on the input and the operation is responsible for finding the types of classifiers (Classification Services), which can be trained using this dataset. For instance, dataset can contain missing values and only those classifiers can be filtered, which are able to deal with missing values. In second operation some other requirements can be specified, for instance the classifier must be updatable. Third operation requires only model capabilities without putting the dataset on the output.

5 Implementation and Use Case Example

The implementation of SODM System is compatible with SOA standards. Each of SODM services is represented using WSDL language. Services are implemented in Java using Weka library [16]. To present functionalities of the SODM system we consider Educational System. One of the problems, which occurs in such systems, is to divide user (students) into priority groups [9]. The priorities of the users can be used to

divide them to early and late registration on courses groups, or to filter students for scholarships awards. Assume that students are described with following three attributes: Average mark form whole studying period (AM), Number of uncompleted courses (NoUC), Number of Awards obtained outside the university (NoA). Educational System collects the past observation of students and their priorities. The goal is to classify the students to priority groups using the knowledge extracted from past observations (training data). Implementing new data mining components in the system can solve the problem but it is much simpler to use Data Mining Services from SODM System. To solve the problem it is necessary to find the type of classifier suitable for the data. To do this Classification Service Manager service can be invoke by putting on the input on the operation representative portion of the data. It is recommended to put the data, which is going to be further used to build the classifier (training data). As a response a specification of Classification Services which corresponds to classifiers types which can be build basing on given on the input dataset are returned. Next, one of the Classification Services can be used to create the model of classifier by invoking buildClassifier operation of these services. This operation takes on the input training datasets (past observations of students and their priorities in the considered example) and returns the key to the created model. The model can be further used to classify objects (unlabeled students) by invoking classifyInstances operation.

6 Final Remarks and Future Works

In this paper we presented basic concepts of SODM System. The system is based on SOA paradigm so the components are fully accessible via ESB. We presented functionalities of Classification and Regression component of SODM System. In future works other data mining components must be developed to create complete service oriented data mining tool. In particular we will apply ensemble classifiers to improve the accuracy of classification.

Acknowledgements

The research presented in this work has been partially supported by the European Union within the European Regional Development Fund program no. POIG.01.03.01-00-008/08.

References

1. Cannataro, M., Congiusta, A., Pugliese, A., Talia, D., Trunfio, P.: Distributed data mining on grids: Services, tools, and applications. *IEEE Transactions on Systems, Man, and Cybernetics* 34(6), 2451–2465 (2004)
2. Fiolet, V., Olejnik, R., Lefait, G., Tournel, B.: Optimal Grid Exploitation Algorithms for Data Mining, pp. 246–252 (2006)
3. Garcia-Laencina, P.J., Sancho-Gomez, J.L., Figuerias-Vidal, A.R.: Pattern Classification with Missing Data: a Review. *Neural Comput. & Applic.* 19, 263–282 (2010)

4. Kuncheva, L.: Combining Pattern Classifiers: Methods And Algorithms. A John Wiley & Sons, Inc., Publication, West Sussex (2004)
5. Kurzyński, M.: Pattern recognition: statistical methods. Oficyna Wydawnicza PWr Wrocław (1997)
6. Marques De Sa, J.P.: Pattern Recognition – Concepts, Methods and Applications. Springer, Oporto University, Portugal (2001)
7. Mitchel, T.M.: Machine Learning. McGraw-Hill Science, New York (1997)
8. Newcomer, E., Lomow, G.: Understanding SOA with Web Services. Addison Wesley Professional, Reading (2004)
9. Prusiewicz, A., Zięba, M.: Services recommendation in systems based on Service Oriented Architecture by applying modified ROCK algorithm. Communications in Computer and Information Science, 226–238 (2010)
10. Rosen, M., Lublinsky, B., Smith, K.T., Balcer, M.J.: Service-Oriented Architecture and Design Strategies. Wiley Publishing, Inc., Chichester (2008)
11. Secretan, J., Georgopoulos, M., Koufakou, A., Cardona, K.: APHID: An architecture for private, high-performance integrated data mining. Future Generation Computer Systems 26, 891–904 (2010)
12. SOA Reference Model Technical Committee. A Reference Model for Service Oriented Architecture, OASIS (2006)
13. Tsoumakas, G., Bassiliades, N., Vlahavas, I.: A knowledge-based web information system for the fusion of distributed classifiers, pp. 271–308. IDEA Group, USA (2004)
14. Tomczak, J., Świątek, J., Brzostowski, K.: Bayesian Classifiers with Incremental Learning for Nonstationary Datastreams. Advances in System Science, 251–261 (2010)
15. WEB SERVICES OVERVIEW,
<http://publib.boulder.ibm.com/infocenter/rtnlhelp/v6r0m0/index.jsp?topic=/com.ibm.etools.webservice.doc/concepts/cws.html>
16. Weka, <http://www.cs.waikato.ac.nz/ml/weka/>
17. Witten, I.H., Frank, E.: Data Mining. Practical Machine Learning Tools and Techniques. Elsevier, San Francisco (2005)

Personalisation in Service-Oriented Systems Using Markov Chain Model and Bayesian Inference

Jakub M. Tomczak and Jerzy Świątek

Institute of Computer Science, Faculty of Computer Science and Management,
Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
{Jakub.Tomczak,Jerzy.Swiatek}@pwr.wroc.pl

Abstract. In the paper a personalization method using Markov model and Bayesian inference is presented. The idea is based on the hypothesis that user's choice of a new decision is influenced by the last made decision. Thus, the user's behaviour could be described by the Markov chain model. The extracted knowledge about users' behaviour is maintained in the transition matrix as probability distribution functions. An estimation of probabilities is made by applying incremental learning algorithm which allows to cope with evolving environments (e.g. preferences). At the end an empirical study is given. The proposed approach is presented on an example of students enrolling to courses. The dataset is partially based on real-life data taken from Wrocław University of Technology and includes evolving users' behaviour.

Keywords: modeling, Markov chain, Bayesian inference, incremental learning.

1 Introduction

Many modern computer networked systems (e.g. e-commerce, web services) are used to provide services to users. Moreover, if the services play a crucial role in the system, such systems could be called *service-oriented systems* (SOSs) [2], [5], [6], [7], [8]. However, in SOSs there are several main problems that have to be concerned [2], [7]: i) user's demand formulation and contract negotiation; ii) user's demand matching with accessible services including aspects such as e.g. knowledge about users' behaviour; iii) service execution on physical machines concerning network quality of service (QoS).

To solve the mentioned problems a process consisting of negotiation, discovery, and execution stages could be proposed. The stage for demand translation and contract negotiation could be based on ontology knowledge representation [10]. Next, in the service discovery there are three main procedures: i) service matching with user's demand (contract), ii) personalisation of services, iii) new service composition if there is no accessible service fulfilling user's demand. Service matching could be made by applying rough set theory [2], and service composition – i.e. using ontologies [8]. Because the personalisation is the main topic of this work, thus it will be discussed in Section 3. The last stage (execution) has to handle final aspects of the whole process and be QoS-aware [5], [6], [7].

This paper consists of following sections. In Section 2 a contribution of proposed approach to the technological innovation is presented. Next, the general problem of

personalisation is described and the solution using Markov chains is outlined. Moreover, a recursive expression for decision making using Bayesian inference is presented. At the end an empirical study is conducted. The presented approach is applied to the real-life problem of students enrolment to courses at Wrocław University of Technology (WTU) and could be used as a new functionality in the existing educational platform (so called *EdukacjaCL*).

2 Contribution to Sustainability

Applying the personalisation method in service-oriented systems aims in increasing quality of user service. It is widely used in the e.g. e-commerce [1], but there is a constant need for new approaches and applications. In this paper an approach of a Markov model as a knowledge representation and Bayesian inference as a reasoning method is proposed. Furthermore, an interesting result of using Markov model and Bayesian inference is that the decision is made due to a recursive procedure (see Section 3.1). Moreover, using knowledge about users gives SOSs a new function of sustainability.

Moreover, an another novelty is using an *incremental learning paradigm* for probabilities estimation [14]. Learning about users' preferences, which evolve in time, is one of the crucial aspects in modern SOSs. And to solve it some adaptive approach has to be applied. Applying incremental learning affects in sustaining model accuracy.

Therefore, in this paper a compact framework using probabilistic model and inference for personalisation problem is proposed. The framework tries to maintain the system's sustainability by making optimal decisions about users' demands and keep an up-to-date knowledge about users.

3 Personalisation in Service-Oriented Systems

Mining knowledge about users in computer systems could lead to increasing quality of user service and profits to service provider(-s), e.g. to overcome so called *information overload* [1], [11], [15]. Therefore, there are different aspects of personalisation, concerning e.g. recommendation, user tutoring, adaptation of layout and content, and so on [15].

However, in this paper the main concern is put on the recommendation task. In other words, on example of an educational platform, during student's enrolment to a course, a sorted list of courses *the best suited* to her/his demands is presented. In the literature there are different approaches to personalisation [1], e.g. individual or collaborative, user or item information, memory- or model-based.

In presented approach it is assumed that user's decision depends on previously made decisions. For example, a student enrols to a new course based on courses she/he has been enrolled in the past. Besides, the decision is rather uncertain because of e.g. lack of information, and that is why a new decision could be described by a probability distribution function. Hence, the Markov chain model [3], [16] seems to be a proper knowledge representation to model users' behaviour because it allows to model a situation that previous decisions affects the current decision. Then, to reason about the maintained knowledge the Bayesian inference is used.

3.1 Problem Statement

As it was mentioned in previous section, a users' behaviour is modelled using Markov chains. Thus, the problem of personalisation could be stated as a classification task.

Let assume that an input in the N^{th} moment is described by a vector of features, $\mathbf{u}_N = \mathbf{U}_1 \times \mathbf{U}_2 \times \dots \times \mathbf{U}_S$ ($\text{card}\{\mathbf{U}_s\} = K_s < \aleph_0$). The sequence of inputs is denoted by $\bar{\mathbf{u}}_N = (\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_N)$. A decision in the N^{th} time step is denoted by $j_N \in \mathbb{M} = \{1, 2, \dots, M\}$. Furthermore, we assume that \mathbf{u}_n , and j_n , for each $n = 1, 2, \dots, N$ are realizations of stochastic processes drawn with distributions $P_n(j_n) \in P_n(\bar{\mathbf{u}}_n | j_n)$. However, in further considerations it is assumed that decisions j_1, j_2, \dots, j_N forms a Markov chain [3] described by an initial (*a priori*) vector of probabilities, $p_1, p_1^{(i)} = P_1(j_1 = i), i=1,2,\dots,M$, and a sequence of transition matrices, $P_N = [p_{N,i,j}], i,j=1\dots M$, $p_{N,i,j} = P_N(j_{N+1} = j | j_N = i)$. Besides, it is assumed that the observations of users are independent, $P_N(\bar{\mathbf{u}}_N | j_N) = \prod_{n=1}^N P_n(\mathbf{u}_n | j_n)$.

Thus, the problem could be stated as follows. Having a new user *the best* suited decision should be made. It could be made by minimizing following risk functional

$$R_N(\bar{\Psi}_N) = \sum_{n=1}^N \mathbf{E}[L(J_n, \Psi_n(\mathbf{X}_n))] \quad (1)$$

where $\bar{\Psi}_N = [\Psi_1 \ \Psi_2 \ \dots \ \Psi_N]$ is a vector of decisions, $i_n = \Psi_n(\mathbf{u}_n)$, $\mathbf{E}[\cdot]$ is an expected value and $L(\cdot, \cdot)$ is a chosen loss function. It is easy to notice [3] that to minimize risk functional (1) it is enough to consider

$$R_n(\Psi_n) = \mathbf{E}[L(J_n, \Psi_n(\mathbf{X}_n))] = \int \sum_{m=1}^M L(m, \Psi_n(\bar{\mathbf{u}}_n)) \cdot p_n^{(m)} \cdot P_n(\bar{\mathbf{u}}_n | m) d\bar{\mathbf{u}}_n \quad (2)$$

where *a priori* distribution in N^{th} moment is $p_n^{(m)} = \sum_{l=1}^M p_{n,m,l} \cdot p_{n-1}^{(l)}$.

Thus, we can propose an optimal decision making algorithm (*Bayesian algorithm*)

$$i_n = \arg \min_{l=1,2,\dots,M} \{r(l, \bar{\mathbf{u}}_n)\} \quad (3)$$

where: $r(l, \bar{\mathbf{u}}_n) = \sum_{m=1}^M L(m, l) \cdot p_n^{(m)} \cdot P_n(\bar{\mathbf{u}}_n | m)$, and for 0-1 loss function:

$$\delta(l, \bar{\mathbf{u}}_n) = p_n^{(l)} \cdot P_n(\bar{\mathbf{u}}_n | l) = \delta_n^{(l)} \quad (4)$$

It could be noticed [3] that using assumption about independence of input observations and that *a priori* distributions depends on recent *a priori* distribution and current transition matrix, the expression for dependent risk functional (5) could be calculated using following recursive procedure:

$$\begin{aligned}\delta_{N+1} &= D_{N+1}(\mathbf{u}_{N+1}) \cdot P_N \cdot \delta_N, \\ \delta_1 &= D_1(\mathbf{u}_1) \cdot p_1\end{aligned}\quad (5)$$

where $\delta_N = [\delta_N^{(1)} \ \delta_N^{(2)} \dots \delta_N^{(M)}]^T$ and

$$D_N(\mathbf{u}_N) = \begin{bmatrix} P_N(\mathbf{u}_N | 1) & 0 & \dots & 0 \\ 0 & P_N(\mathbf{u}_N | 2) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & P_N(\mathbf{u}_N | M) \end{bmatrix} \quad (6)$$

Hence, to make a decision for a given input in N^{th} time step it is enough to have matrices $D_N(\mathbf{u}_N)$, P_N , and the vector δ_{N-1} . It is important due to the learning stage.

3.2 General Methodology

However, from the computational point of view it is important to have as small number of features as possible with smallest loss of information. Therefore, the number of inputs could be decreased to minimum. For example, a student could be described by a mean value of grades, number of courses he attended. Then students could be clustered into groups and each group contains students who have similar descriptions. One group is so called *context* [4], [9].

Furthermore, in real situations the probability distributions are unknown. Therefore, a learning stage is needed. During the learning process the distributions are estimated based on training sequence, (\mathbf{u}_{n,j_n}) , $n = 1, 2, \dots, N$. It is worth to notice, that at each n there could be more than one observation, $(\mathbf{u}_{n,k,j_{n,k}})$, $k=1,2,\dots,K$. However, in real environments the non-stationarity occurs which means that the estimation cannot be made using all observations, e.g. users' preferences which evolve in time [15].

Thus, following general methodology could be propose:

1. Divide users descriptions (demands) into clusters (each cluster is called a context) using some chosen clustering algorithm (e.g. *k-Means*).
2. *Learning Stage*:
 - i.) At the beginning estimate $p_1, D_1(\cdot)$
 - ii.) At each learning step $N > 1$ estimate $D_N(\cdot)$, P_N , and δ_{N-1} .
3. *Classification stage*:
 - i.) For given user's demand find an according context (e.g. using *1-Nearest Neighbour* classifier).
 - ii.) For given user's context make a decision using (5).

3.3 Incremental Learning Algorithm

As it was mentioned, in many real-life situation the non-stationarity occurs. Therefore, an adaptive estimation methods have to be proposed. Such adaptation to changes could be made by applying *incremental* learning algorithms [14].

In presented approach a following method of estimation is proposed. Let us assume that at each learning step there are K observations which come in a data stream (sequence). It means that k^{th} observation occurs before $(k+1)^{\text{th}}$ observation. Then, the

probability distribution is estimated using a frequency matrices (matrices of the same sizes as $D_N(\cdot)$, P_N , and δ_{N-1}) but consisting frequency of occurrence of responding input or decision. Let denote such matrix as $V_{D,N}$, $V_{P,N}$, $V_{\delta,N-1}$. Then, the learning algorithm could be as follows¹:

1. Take a new observation and initiate all matrices.
2. Update the appropriate frequency matrix by a new k^{th} observation,

$$V_N^{(k)} = a_N \cdot V_{N-1} + V_N^{(k-1)} + W_k$$

where W_k is a matrix with ones in proper places for observation's decision and context (determining row and column), and zeros – otherwise, a_N $[0,1]$ is a forgetting factor.

3. Estimate probability distribution using appropriate frequency matrix. If there is a new observation, then go to 1.

The key issue is to fix a proper value of the forgetting factor because it affects the estimation. It can have a different value at each learning step or be constant.

4 Experimental Study

Presented approach is checked on the example of the educational platform at WTU. The platform is dedicated to students service and enables e.g. sending applications, signing up for courses, checking past grades, and so on. Therefore, the educational platform could be seen as a SOS in which different services could be distinguished.

However, in this work the proposed approach is used in the enrolment service. At WTU there are different services for enrolment: faculty enrolment, specialisation enrolment, sport enrolment, and foreign languages enrolment. Each of mentioned enrolment needs reading about tens or even hundreds of course descriptions which is a big waste of time both for student and system. In the experiment on the example of a enrolment for foreign language it is shown how proposed approach could be used.

4.1 Experiment Details and Results

In the experiment the state in the Markov chain is associated with the language and its level: *English A1*, *English A2*, *English B1*, *English B2*, *English C*, *German A1*, *German A2*, *German B1*, *German B2*, *German C*, *nothing*. At WTU there are more languages available (e.g. Korean, Japanese, Czech, Italian, Swedish) but for transparency of the experiment it was limited to 11. The *nothing* means that student is not signed up for any course.

Moreover, it is assumed that each student is described by a following vector: *number of all courses*, *number of exercises*, *number of laboratories*, *number of lectures*, *mean of grades*, *variance of grades*, *mean of grades from lectures*, *mean of grades from exercises*, *mean of grades from laboratories*, *number of fails (grade F)*, *number of excellents (grade A)*. To generate students a real logs from the educational platform was used. This dataset was used also in the previous works [12], [13].

¹ Because the algorithm is the same for all matrices, therefore indexes by V are skipped.

Then the problem is as follows: ***Propose a student the most appropriate language and the level if recently she/he was enrolled to a course X.***

The experiment was conducted for 2 semester (one semester – one learning stage). In the semester one student signs up for a course after another.

The methodology of the experiment was following:

1. Using the real dataset a context was established using *k-Means* clustering algorithm. (There are 3 clusters).
2. Student is generated due to the appropriate probability distribution² and her/his description is mapped with the context id by using *1-NN* classifier.
3. Initial state and next states are generated due to (5).
4. Incremental learning algorithm with a forgetting factor a is applied.

It was assumed that at the first learning stage (first semester) $a_1 = 0$.

Following methods were compared (for 100 students during a semester, 2000 students, and 5000 students):

- Markov chain (MC) model with Bayesian inference (5) with $a_2 \in \{0, 0.1, 0.25, 0.5\}$.
- Bayesian model (no assumption about Markov chain, *Bayes*) with Bayesian inference with $a_2 \in \{0, 0.1, 0.25, 0.5\}$.
- Random – list of states is given randomly.

Additionally, beside classification accuracy another performance index was used, called *position index*. This criterion calculates the difference of real state in the sorted due the probabilities list of decisions from the first position.

Table 1. Classification accuracy (mean value and standard deviation) for compared models

	Mean			Std		
	100	2000	5000	100	2000	5000
MC 0.0	0.23	0.283	0.285	0.043	0.005	0.004
MC 0.1	0.229	0.283	0.285	0.042	0.005	0.004
MC 0.25	0.229	0.283	0.285	0.042	0.005	0.004
MC 0.5	0.228	0.282	0.285	0.04	0.005	0.004
Bayes 0.0	0.245	0.281	0.281	0.046	0.009	0.0055
Bayes 0.1	0.245	0.281	0.281	0.047	0.009	0.0055
Bayes 0.25	0.244	0.281	0.281	0.047	0.009	0.0055
Bayes 0.5	0.244	0.281	0.281	0.047	0.009	0.0055
Random	0.095	0.093	0.090	0.015	0.004	0.001

The experiment was conducted on 10 generated datasets (decisions about enrolments) and the results are a mean values. Classification accuracy and position index are calculated as a mean of two semesters. The results are shown in the tables 1 (*classification accuracy*) and 2 (*position index* – the lower the value the better).

4.2 Discussion

First of all it has to be said that analysing classification accuracy for Markov model and Bayesian model no statistical difference could be noticed. In the situation with

² Features *number of fails* and *number of excellents* were drawn from discrete distributions and all other – Gaussian.

Table 2. Position index (mean value and mean value of standard deviation) for compared models

	Mean value of position index			Mean value of std of position index		
	100	2000	5000	100	2000	5000
MC 0.0	2,91	2,46	2,42	0,713	0,597	0,583
MC 0.1	2,85	2,45	2,42	0,662	0,592	0,581
MC 0.25	2,86	2,45	2,42	0,660	0,592	0,581
MC 0.5	2,87	2,43	2,42	0,665	0,591	0,581
Bayes 0.0	2,81	2,53	2,53	0,703	0,745	0,756
Bayes 0.1	2,78	2,53	2,53	0,636	0,741	0,754
Bayes 0.25	2,78	2,53	2,53	0,636	0,741	0,754
Bayes 0.5	2,76	2,53	2,53	0,639	0,740	0,754
Random	4,95	4,97	6,02	0,230	0,277	0,356

100 students in semester one and two the Markov model performs worst because of its bigger size of matrices. However, in case of 2000 and 5000 students Markov model gave slightly better results. On the other hand, analysing position index Markov model was a little bit better than Bayesian model and the mean value of std shows that Markov model behaves in more stable way (difference of about 0.15).

Furthermore, applying Bayesian inference allows to propose around 1/3 of students proper decision or, in average, proper decision at 2nd or 3rd position in the list. In comparison to random method it is clear that Bayesian approach outperforms *random* personalisation. Moreover, the usage of the incremental algorithm gives slight improvement in the quality of classification accuracy and position index.

Concluding, presented experiment is conducted only on two semesters. Probably, carrying out the experiment on 10 semesters should show that applying Markov chain model is more appropriate.

5 Final Remarks

In this paper the general approach using Markov chain model and Bayesian inference for personalisation was proposed. The presented method was evaluated on the partially real-life data take from the educational platform. The problem was formally stated and the general methodology was proposed.

In the future more attention should be paid in developing the whole system with specified, co-working modules. Moreover, given methodology should be applied in the existing educational platform as the additional feature.

Furthermore, the Markov model should be compared with other methods, e.g. using rules-based knowledge representation, or ensemble classifiers. And the higher order Markov chains are supposed to be considered.

Besides, the research ought to be conducted on bigger amount of data, including logs of users containing whole history, e.g. history of all enrolments.

Acknowledgments. The research presented in this paper has been partially supported by the European Union within the European Regional Development Fund program no. POIG.01.03.01-00-008/08.

References

1. Anand, S.S., Mobasher, B.: Intelligent Techniques for Web Personalization. In: Mobasher, B., Anand, S.S. (eds.) ITWP 2003. LNCS (LNAI), vol. 3169, pp. 1–36. Springer, Heidelberg (2005)
2. Brzostowski, K., Tomczak, J.M., Nikliborc, M.: A Rough Set-Based Algorithm to Match Services Described by Functional and Non-Functional Features. In: Grzech, A., et al. (eds.) Networks and Networks' Services, Ofic. Wyd. PWr, Wrocław, pp. 27–38 (2010)
3. Bubnicki, Z.: Pattern Recognition Algorithms for Simple Markov Chains. In: Problemy informacji i sterowania, Ofic. Wyd. PWr, Wrocław, pp. 3–18 (1972) (in Polish)
4. Chen, N., Chen, A.: Integrating Context-Aware Computing in Decision Support System. In: Proc. of the Int. ME & CS 2010, Hong Kong, March 17–19, vol. I (2010)
5. Grzech, A., Rygielski, P.: Translations of Service Level Agreement in Systems Based on Service Oriented Architecture. In: Setchi, R., Jordanov, I., Howlett, R.J., Jain, L.C. (eds.) KES 2010. LNCS, vol. 6277, pp. 523–532. Springer, Heidelberg (2010)
6. Grzech, A., Rygielski, P., Świątek, P.: QoS-aware infrastructure resources allocation in systems based on service-oriented architecture paradigm. In: HET - NETs, pp. 35–47 (2010)
7. Grzech, A., Świątek, P.: Modeling and Optimization of Complex Services in Service-Based Systems. *Cybernetics and Systems: An International Journal* 40, 706–723 (2009)
8. Juszczyszyn, K., Stelmach, P., Grzelak, T.: A Method for the Composition of Semantically Described Web Services. In: Grzech, A., et al. (eds.) Networks and Networks' Services, Ofic. Wyd. PWr, Wrocław, pp. 27–38 (2010)
9. Palmisano, C., Tuzhilin, A., Gorgoglione, M.: Using Context to Improve Predictive Modeling of Customers in Personalization Applications. *IEEE Trans. on Knowledge and Data Engineering* 20(11), 1535–1549 (2008)
10. Pastuszko, M., Kryza, B., Ślota, R., Kitowski, J.: Processing and negotiation of natural language based contracts for Virtual Organizations. In: Proc. of Cracow 2009 Grid Workshop, ACC CYFRONET AGH, Kraków, pp. 104–111 (2010)
11. Pierrakos, D., Palioras, G., Papatheodorou, C., Spyropoulos, C.D.: Web Usage Mining as a Tool for Personalization: A Survey. *User Mod. & User-Ad. Inter.* 13, 311–372 (2003)
12. Prusiewicz, A., Zięba, M.: Services Recommendation in Systems Based on Service Oriented Architecture by Applying Modified ROCK Algorithm. In: Zavoral, F., Yaghob, J., Pichappan, P., El-Qawasmeh, E. (eds.) NDT 2010. CCIS, vol. 88, pp. 226–238. Springer, Heidelberg (2010)
13. Sobecki, J., Tomczak, J.M.: Student courses recommendation using Ant Colony Optimization. In: Nguyen, N.T., Le, M.T., Świątek, J. (eds.) Intelligent Information and Database Systems. LNCS (LNAI), vol. 5991, pp. 124–133. Springer, Heidelberg (2010)
14. Tomczak, J.M., Świątek, J., Brzostowski, K.: Bayesian Classifiers with Incremental Learning for Nonstationary Datastreams. In: Grzech, A., Świątek, P., Drapała, J. (eds.) Advances in Systems Science, pp. 251–260 EXIT, Warszawa (2010)
15. Webb, G.I., Pazzani, M.J., Billsus, D.: Machine Learning for User Modeling. *User Modeling and User-Adapted Interaction* 11, 19–29 (2001)
16. Zhu, J., Hong, J., Hughes, J.G.: Using Markov Chains for Link Prediction in Adaptive Web Sites. In: Bustard, D.W., Liu, W., Sterritt, R. (eds.) Soft-Ware 2002. LNCS, vol. 2311, pp. 55–66. Springer, Heidelberg (2002)

Automatic Extraction of Document Topics

Luís Teixeira¹, Gabriel Lopes², and Rita A. Ribeiro¹

¹ CA3-Uninova, FCT, Universidade Nova de Lisboa 2829-516 Caparica, Portugal

{lst,rar}@uninova.pt

² DI-FCT/UNL, 2829-516 Caparica, Portugal

{gpl}@fct.unl.pt

Abstract. A keyword or topic for a document is a word or multi-word (sequence of 2 or more words) that summarizes in itself part of that document content. In this paper we compare several statistics-based language independent methodologies to automatically extract keywords. We rank words, multi-words, and word prefixes (with fixed length: 5 characters), by using several similarity measures (some widely known and some newly coined) and evaluate the results obtained as well as the agreement between evaluators. Portuguese, English and Czech were the languages experimented.

Keywords: Document topics, words, multi-words, prefixes, automatic extraction, suffix arrays.

1 Introduction

A topic or a keyword of a document is any word or multi-word (taken as a sequence of two or more words, expressing clear cut concepts) that summarizes by itself part of the content of that document belonging to a collection of documents.

The Extraction of topics (or keywords) is useful in automatic construction of ontologies, document summarization, clustering and classification, and to enable easier and more effective access to relevant information in Information Retrieval. To measure the relevance of a term (word or multi-word) in a document one must take into account the frequency of that term in that document and in the rest of document collection. Desirably, that term should not appear or should be rare in documents focusing on other subject matters.

Tf-Idf, phi-square, mutual information and variance are measures often used to deal with term relevance in documents and document collections ([16] and [1]). In this paper we use those measures (and newly coined variants of them) to extract both single-words and multi-words as key-terms, and compare the results obtained. Additionally, we identify relevant prefixes (with 5 characters length) in order to deal with morphologically rich languages. As no one is able to evaluate prefixes as relevant or non-relevant, we had to project (bubble) prefix relevance into words and multi-words and created, for this purpose, a new operator (bubble) and new relevance measures) to enable the bubbling of prefix relevance, first into corresponding words, and later in multi-words. Simultaneously, we improve discussion started in [1] and continued in [10] and arrive at different conclusions, namely that results obtained by using tf-idf, phi-square and newly derived measures are better than results obtained by using mutual information or variance and derived measures.

In section 2 we describe how our work contributes to sustainability; related work is summarized in section 3. In section 4 and 5 the measures used are defined; experiments done are described in section 6 and the results obtained are presented in section 7. In section 8 we draw the conclusions on this paper.

2 Contribution to Sustainability

This work impacts on sustainability when easy and intelligent access to large document collections is a stake. Our computations use suffix arrays as an adequate data structure and contribute to decrease computing time and power consumption, thus providing new ways to power saving on high performance search centers.

3 Related Work

In [2], [3], and [4] authors propose systems to extract noun phrases and keywords using language depend tools such as stop-words removing, lemmatization, part-of-speech tagging and syntactic pattern recognition. As it will be seen, our work diverges from those ones as it is clearly language independent.

The work in [5] and [6], for multi-word term extraction, rely on predefined linguistic rules and templates to be able to identify certain type of entities in text documents, making them language dependent. In this area, the method proposed in [10] for extracting multi-words, requiring no language knowledge, will be used for extracting multi-words in 3 languages (EN, PT and CZ), as reported in this paper.

In [7] the extraction of Key-words from news data is approached. This is a non-language independent work. A supervised approach for extracting keywords is proposed in [8], using lexical chains built from the WordNet ontology [9], a tool not available for all languages. In [1], the paper that motivated our work, a Key-term extractor (multi-words) is presented together with a metric, the LeastRvar. However, single words are ignored. From the same authors, in [10], the extraction of single and multi-words as key-terms is worked out. However, a share quota for most relevant single and multi-words is predefined, assuming multi-words as better key-terms. In our work, words, multi-words and prefixes are treated identically, with no predefined preferences. Results obtained support this other vision and show that tf-idf and Phi-square-based measures outperform Rvar and Mutual Information based metrics.

4 Measures Used

In this section, for the purpose of completeness, some well-known measures used in this work are presented, as well as those newly coined measures we had to create.

4.1 Known Measures Used

Tf-Idf Metric. Tf-Idf (Term frequency-Inverse document frequency) [1] is a statistical metric often used in information retrieval and text mining. Usually, it is used to evaluate how important a word is to a document in a corpus. The importance increases proportionally to the number of times a prefix/word/multiword appears in the

document but it is offset by its frequency in the corpus. It should be noticed that we use a probability, $p(W, d_j)$, in equation (1), defined in equation (2), instead of using the usual term frequency factor.

$$\text{Tf-Idf}(W, d_j) = p(W, d_j) * \text{Idf}(W, d_j). \quad (1)$$

$$p(W, d_j) = f(W, d_j) / N_{d_j}. \quad (2)$$

$$\text{Idf}(W, d_j) = \log(||D|| / ||\{d_i : W \in d_i\}||). \quad (3)$$

Where $f(W, d_j)$ denotes the frequency of prefix/word/multiword W in document d_j and N_{d_j} stands for the number of words of d_j ; $||D||$ is the number of documents of the corpus. So, $\text{Tf-Idf}(W, d_j)$ will give a measure of the importance of W within the particular document d_j . By the structure of term Idf we can see that it privileges prefixes, multi-words and single words occurring in fewer documents.

Rvar and LeastRvar, two measures based on variance, were first presented in [1], with the aim of measuring the relevance of multi-words extracted automatically [16], and are formulated as follows:

$$\text{Rvar}(W) = (1 / ||D||) * \sum (p(W, d_i) - p(W, .))^2. \quad (4)$$

where $p(W, d_j)$ is defined in (2) and $p(W, .)$ is the median probability of word W taking into account all documents. Being MW a multi-word made of word sequence $(W_1 \dots W_n)$, LeastRvar is determined as the minimum of $\text{Rvar}()$ applied to the leftmost and rightmost words of MW.

$$\text{LeastRvar}(\text{MW}_i) = \min (\text{Rvar}(W_1), \text{Rvar}(W_2)). \quad (5)$$

Phi Square Metric. The Phi Square [12] is a variant of the known measure Chi-Square, allowing a normalization of the results obtained with the Chi Square, and is given by the following expression:

$$\phi^2 = (N \cdot (AD - CB)^2 / (A+C)(B+D)(A+B)(C+D)) / M. \quad (6)$$

Where M is the total number of terms present in the corpus (the sum of terms from the documents that belong to the collection). And where A is the number of times term t occurs in document d ; B the number of times that term t occurs in the other documents of the corpus; C stands for the number of terms of the document d subtracted by the amount of times term t occurs in document d ; D is the number of times that neither document d or term t occur (i.e. $D = N - A - B - C$); and N the total number of documents.

Mutual Information. This measure [15] is widely used in language modulation, and its intent is to identify associations between randomly selected terms and in that point determine the dependence that those terms have among them. This measure presented poor results.

4.2 New Measures Used

It was important to have all measures treated the same way. So, if operator “least” was applied to Rvar [1], it should be applied to any other measure used to rank relevance of words, multi-words and prefixes. So, in this section, we describe the newly

created measures based on operators “Least” and “Bubbled”. In the following, equations consider that MT stands for any of the used measures on this work (Tf-Idf, Rvar, Phi-square or ϕ^2 , and Mutual Information or MI), P a Prefix, W a word, and MW a multi-word made of word sequence ($W_1 \dots W_n$).

Least Operator. This operator is the same used in the measure LeastRvar, adapted to work with words alone, where we assume that the leftmost and rightmost words of a single word coincide with the word itself.

$$\text{Least_MT}(W) = \text{MT}(W). \quad (7)$$

$$\text{Least_MT}(MW) = \text{Min}(\text{MT}(W_1), \text{MT}(W_n)). \quad (8)$$

Bubbled Operator. Another problem we needed to solve was the propagation of the relevance of each Prefix to words having it as a prefix.

$$\text{Bubbled_MT}(W) = \text{MT}(P). \quad (9)$$

Having the operators defined we can now present the formulation for the new metrics used.

$$\text{Least_Bubbled_MT}(W) = \text{Bubbled_MT}(P). \quad (10)$$

$$\text{Least_Bubbled_MT}(MW) = \text{Min}(\text{Bubbled_MT}(W_1), \text{Bubbled_MT}(W_n)). \quad (11)$$

As in [10] the median of word length in characters was used to better rank words and multi-words, we consider two additional operators: LM for “Least_Median” and LBM, for “Least_Bubbled_Median” defined in (12) and (13), where T represents a term (word or multi-word).

$$\text{LM MT}(T) = \text{Least_MT}(T) * \text{Median}(T). \quad (12)$$

$$\text{LBM MT}(T) = \text{Least_Bubbled_MT}(T) * \text{Median}(T). \quad (13)$$

5 Experiments

We worked with a collection of parallel texts, common for the three languages experimented, Portuguese, English and Czech, from European legislation in force (<http://eur-lex.europa.eu/>). The total number of terms for these collections was of 109449 for Portuguese, 100890 for English and 120787 for Czech.

Multi-words were extracted using LocalMax algorithm [10] as implemented in [13]. SuffixArrays [14] were used for word extraction and for multi-words, words and prefixes counting.

We worked with single words having a minimum length of six characters (this parameter is changeable) and filtered multi-words (with words of any length) removing those containing punctuation marks, numbers and other symbols. Results presented in tables bellow are based on the evaluation of one of the two evaluators, the most critic one. Table 1 shows the top best ranked terms extracted from 3 parallel documents. Tables 2 and 3 show, for the subset of measures used, that were directly evaluated, the average precision obtained for the three languages for one Evaluator.

Evaluators were asked to evaluate 25 best ranked terms for each one of the six measures in those tables. The evaluation assigned a classification (good topic descriptor (G), near good topic descriptor (NG), bad topic descriptor (B), unknown (U), and not evaluated (NE). Last classification (NE) was required because evaluation was indirectly propagated for the rest of measures that were not directly evaluated. K-statistics, used to measure the degree of agreement between evaluators, is shown in table 3, for measures specifically evaluated. Table 4 shows average precision for the N top ranked terms for best evaluated measures with N equal to 5, 10 and 20. In tables 2, 3 and 4, L was used for Least Operator, LM for Least Median Operator, LBM for Least Bubbled Median operator.

6 Results

Some of the results obtained are presented in the following tables.

Table 1. First five terms extracted, ranked accordingly using the measure Phi-Square for all languages for a document in the corpus

Portuguese	Czech	English
multilinguismo (G)	Mnohojazyčnost (G)	Multilingualism (G)
alto nível sobre o multilinguismo (NG) nomeados a título (B)	Podskupiny (NG)	group on multilingualism (G)
domínio do multilinguismo (G)	Mnohojazyčnosti (G)	high level group on multilingualism (NG)
composto por oito (B)	Skupiny (NG)	members of the group (B)
	yyské úrovní pro mnohojazyčnost (G)	sub-groups (B)

Table 2. Average Precision for 5 best ranked terms for Evaluator 1 and all Languages

	ϕ^2	L tfIfd	LM Rvar	LM MI	LBM ϕ^2	LBM Rvar
Portuguese	0,723	0,6389	0,463	0,424	0,6222	0,517
English	0,844	0,785	0,472	0,472	0,800	0,524
Czech	0,700	0,750	0,450	0,450	0,550	0,500

Table 3. K-statistics for the two evaluators

		ϕ^2	L tfIfd	L M Rvar	LM MI	LBM ϕ^2	LBM Rvar
K- Statistics	Portuguese	0.552	0.6324	0.11	0.0196	0.635	0.2152
	English	0.7275	0.4375	0.2665	0.2584	0.5786	0.3478

Table 4. Average precision for the best ranked 5,10and 20 terms, for CZ, EN and PT using best applied measures

	Czech			English			Portuguese		
	P(5)	P(10)	P(20)	P(5)	P(10)	P(20)	P(5)	P(10)	P(20)
LM tfidf	0.70	0.65	0.59	0.81	0.78	0.66	0.68	0.63	0.64
LB tfidf	0.80	0.68	0.65	0.85	0.66	0.65	0.86	0.71	0.65
LM ϕ^2	0.70	0.60	0.58	0.87	0.78	0.70	0.61	0.64	0.59
L ϕ^2	0.70	0.60	0.58	0.83	0.76	0.69	0.68	0.64	0.59
LBM tfidf	0.65	0.68	0.66	0.82	0.69	0.62	0.83	0.70	0.68
tfidf	0.90	0.86	0.66	0.84	0.74	0.67	0.69	0.70	0.66

7 Discussion

As shown in table 3, agreement between evaluators was higher for the specifically evaluated measures Phi Square, Least Tf-Idf, and Least Bubbled Median Phi-Square. Propagated evaluation to other Tf-idf and ϕ^2 based measures also showed equivalent agreement results. MI and ϕ^2 based measures obtained poorer agreement.

Contradicting the point of view presented at [1], we may say that Tf-Idf is a good measure for selecting key-terms. Moreover the terms extracted by both Tf-Idf and Phi-square, or any of its new variants, show better results than the ones obtained by Rvar, or any of its variants, which were considered better than Tf-df in [1].

Rvar and Mutual Information alone were not capable of adequately ranking terms. Only the usage of variants of these measures, applying Least, Bubble and Median operators, improved their results and enabled a selection of best first terms. Otherwise first 200 or 400 terms would be equally ranked.

Evaluated results in table 4 for Portuguese and Czech, two highly inflected languages, are equivalent. Average precision for English is approximately 10% higher than the values obtained for Portuguese or Czech. Best precision results are obtained with different ranking measures for the evaluation of the N best selected key-terms. Tf-Idf alone produces best results for Czech. Least Bubbled Median Tf-Idf is the best for 20 higher ranked key-terms in Portuguese and Czech. Least Median Phi Square and Least Median Tf-Idf works better for English while Least Bubbled Tf-Idf produces better results for Portuguese. Results from variants of Rvar and Mutual Information were always below 55% for all ranges of terms selected (table 2).

Bubbled variants showed rather interesting results for the three languages, especially for Portuguese and Czech. Least and Least Median operators enabled best results for English.

8 Conclusions

Instead of being dependent on specific languages, structured data or domain, we try to approach the key-term extraction problem (of words and multi-words) from a more

general and language independent perspective and make no distinctions between words and multi-words, as both kinds of entities pass the same kind of sieve to be ranked as adequate topic descriptors. Also it can be said that the extraction of prefixes (for dealing with highly inflected languages as is Czech and, to a lower degree, Portuguese) and propagating their relevance into words and multi-words, apart from being one of the main innovations presented, enabled high precision (and recall, not shown) values for the top 20 best ranked topic describing terms extracted.

Also the usage of Suffix Arrays has proved to be very efficient and fast in the extraction of words and prefixes from it, also made viable in a more effective way the counting the occurrences of the words, multi-words and Prefixes within the corpus.

References

1. da Silva, J.F., Lopes, G.P.: A Document Descriptor Extractor Based on Relevant Expressions. In: Lopes, L.S., Lau, N., Mariano, P., Rocha, L.M. (eds.) EPIA 2009. LNCS, vol. 5816, pp. 646–657. Springer, Heidelberg (2009)
2. Cigarrán, J.M., Peas, A., Gonzalo, J., Verdejo, F.: Automatic selection of noun phrases as document descriptors in an FCA-based information retrieval system. In: Ganter, B., Godin, R. (eds.) ICFCA 2005. LNCS (LNAI), vol. 3403, pp. 49–63. Springer, Heidelberg (2005)
3. Liu, F., Pennell, D., Liu, F., Liu, Y.: Unsupervised approaches for automatic keyword extraction using meeting transcripts. In: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics Boulder, Boulder, Colorado, May 31-June 05 (2009)
4. Hulth, A.: Enhancing linguistically oriented automatic keyword extraction. In: Proceedings of Human Language Technology-North American Association for Computational Linguistics 2004 conference, May 02-07, pp. 17–20. Association for Computational Linguistics, Boston (2004)
5. Yangarber, R., Grishman, R.: Machine Learning of Extraction Patterns from Unannotated Corpora: Position Statement. In: Workshop on Machine Learning for Information Extraction. Held in conjunction with the 14th European Conference on Artificial Intelligence (ECAI), August 21. Humboldt University, Berlin (2000)
6. Christian, J.: Spotting and Discovering Terms through Natural Language Processing. MIT Press, Cambridge (2001)
7. Martínez-Fernández, J.L., García-Serrano, A., Martínez, P., Villena, J.: Automatic Keyword Extraction for News Finder. In: Nürnberger, A., Detyniecki, M. (eds.) AMR 2003. LNCS (LNAI), vol. 3094, pp. 99–119. Springer, Heidelberg (2004)
8. Ercan, G., Cicikli, I.: Using lexical chains for keyword extraction. *Information Processing and Management: an International Journal* archive 43(6), 1705–1714 (2007)
9. Miller, G.A.: The science of words. Scientific American Library, New York (1991)
10. de Silva, J.F., Lopes, G.P.: Towards Automatic Building of Document Keywords. In: The 23rd International Conference on Computational Linguistics, COLING 2010, Pequim (2010)
11. de Silva, J.F., Dias, G., Guilloré, S., et al.: Using LocalMaxs Algorithm for the Extraction of Contiguous and Non-contiguous Multiword Lexical Units. In: 9th Portuguese Conference on Artificial Intelligence Evora, September 21-24 (1999)
12. Everitt, B.S.: The Cambridge Dictionary of Statistics, CUP (2002)

13. Multi-Word Extractor,
<http://hlt.di.fct.unl.pt/luis/multiwords/index.html>
14. Suffix arrays, <http://www.cs.dartmouth.edu/~doug/sarray/>
15. Manning, C.D., Raghavan, P., Schütze, H.: An Introduction to Information Retrieval. Cambridge University Press, Cambridge (2008)
16. Sebastiani, F.: Machine Learning in Automated Text Categorization. ACM Computing Surveys 34(1), 1–47 (2002)

Adaptive Imitation Scheme for Memetic Algorithms

Ehsan Shahamatnia¹, Ramin Ayanzadeh², Rita A. Ribeiro¹, and Saeid Setayeshi³

¹ UNINOVA-CA3, UNL-FCT Campus, 2829-516 Caparica, Portugal

E.Shahamatnia@fct.unl.pt, rar@uninova.pt

² Islamic Azad University, Science and Research Campus, Tehran, Iran

ayanzadeh@srbiau.ac.ir

³ Amirkabir University of Technology

setayesh@aut.ac.ir

Abstract. Memetic algorithm, as a hybrid strategy, is an intelligent optimization method in problem solving. These algorithms are similar in nature to genetic algorithms as they follow evolutionary strategies, but they also incorporate a refinement phase during which they learn about the problem and search space. The efficiency of these algorithms depends on the nature and architecture of the imitation operator used. In this paper a novel adaptive memetic algorithm has been developed in which the influence factor of environment on the learning abilities of each individual is set adaptively. This translates into a level of autonomous behavior, after a while that individuals gain some experience. Simulation results on benchmark function proved that this adaptive approach can increase the quality of the results and decrease the computation time simultaneously. The adaptive memetic algorithm proposed in this paper also shows better stability when compared with the classic memetic algorithm.

Keywords: Optimization, Evolutionary computing, Memetic algorithm, Imitation operator, Adaptive imitation.

1 Introduction and Related Works

Increasingly, complex systems in different domains raise challenging problems which cannot efficiently be solved with conventional methods. The quest for a solution to these kinds of problems has led researches to use soft computing techniques, by which they can obtain near optimal solutions [5,6]. Genetic algorithm (GA) is one of the earliest and most renowned metaheuristics successfully applied on many real world problems [13-16]. Genetic algorithms, like many other metaheuristics, such as particle swarm optimization, explore large areas of search space and locate local minima in early iterations but slack off in trying to find the global optimum [1]; this behavior is demonstrated in Fig. 1. Another major problem these metaheuristics face is their instability. Due to their stochastic nature they may produce quite different results in different runs of algorithm [5]. Among many contributions made to overcome these problems and aiming to improve their performance, another family of metaheuristics, called memetic algorithms, has attracted many researchers. A memetic algorithm, henceforth called MA, is a hybridization of a global optimization technique with a cultural evolutionary stage which is responsible for local refinements [17,26]. It is

reported in the literature that the marriage between global and local search is almost always beneficial [1-3,18,23-25]. Memetic algorithms are also known in the literature as Hybrid Evolutionary Algorithms, Hybrid GAs, Baldwinian or Lamarckian Evolutionary Algorithms, Cultural Algorithms or Genetic Local Searchers.

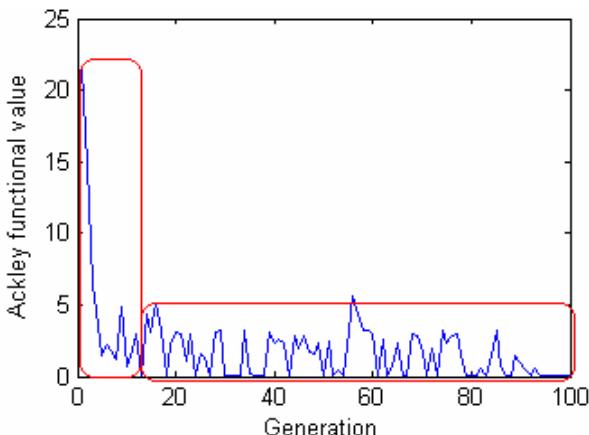


Fig. 1. Sample performance of a genetic algorithm in minimizing a test function

As biologists put it, a gene is the unit of heredity in a living organism via which the physiological characteristics such as eye and hair color is passed from parents to offsprings. First used in 1976 by Richard Dawkins [4], memes are cultural analogues to genes, in that they are responsible for transmitting the behavioral characteristics such as beliefs and traditions [2]. Biologists believe that genes in a chromosome are (normally) intact during their life span, while psychologists argue that memes are from a more dynamic nature, and they can improve and change under the influence of the environment they are exposed to. Actually the cornerstone of memetic algorithms is that individuals can improve their fitness by means of imitation.

To implement this concept in the computational framework, several approaches have been suggested such as multi-meme MA [17,19], coevolving and self generation MA [20] and multi-agent MA [21], but the mainstream is to conduct a type of local search around each possible solution's neighborhood within a predefined radius [5-7]. In the rest of this paper we will consider GA based memetic algorithm but the results can be extended to the most MAs as well. Regarding the interaction of genes and memes in the memetic algorithm evolution process, two main strategies can be witnessed; in memetic algorithms based on Lamarckian theory, the results obtained from imitation operator overwrite the genes. In other words, individuals are strongly affected by the environment and the change in their aptitude is disseminated via the alteration in their genes. In Baldwinian memetic algorithms genes and memes are stored separately [3] as illustrated in figure 2. In this strategy, genes are utilized in the reproduction and memes are utilized in the process of choosing the chromosomes. Hence, among chromosomes with similar fitness, those who have better solutions in their neighborhood have higher chance of reproduction [3].

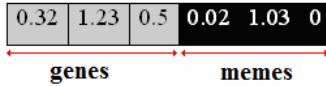


Fig. 2. Sample chromosome in Baldwinian MA

It is well established that pure genetic algorithms cannot easily fine tune the solutions in a complex search space [1], whereas if they are coupled with an individual learning procedure capable of performing local refinements the results will be promising [8,9,22]. Benefiting from both Darwinian evolution and domain specific heuristics, memetic algorithms not only balance the generality and problem specificity, but also render a good methodology to trade off between the exploration capacities of underlying GA and the exploitation capabilities of local search heuristic. In the other hand this translates into more computations, more time and loss of diversity within population [10]. As an example, with a classic hill climbing algorithm, for a chromosome with n genes if each gene can take 3 different values, then in each iteration 3^n different neighbors must be checked. To ameliorate these drawbacks, we first propose an adaptive imitation scheme and then we study the performance of variations of hill climbing search strategy in conjunction with the adaptive imitation.

Beside their strengths, memetic algorithms also have their limitations. This paper addresses two major issues related to memetic algorithm design and performance. First issue is designing an appropriate local search scheme for the problem at hand. For some problems it is very difficult to define a neighborhood mechanism for a point in the search space. Furthermore, local search strategies usually bear high computational time [11]. Second issue is setting the imitation operator parameters. In this paper first we propose an adaptive scheme to define the intensity of imitation operator. Then the effect of this adaptive scheme has been studied on different variations of hill climbing local search.

The rest of the paper is organized as follows. In Section 2, we discuss the technological innovation of this paper for sustainability. In section 3, the proposed adaptive imitation operator is evaluated and different local search strategies are studied. The simulation results are also discussed in this section. Finally, Section 4 concludes this paper.

2 Technological Innovation for Sustainability

This work impacts on sustainability when we are challenged by finding optimal solutions for increasingly complex systems (mostly in terms of dimension of the search space) in different domains, which cannot be efficiently solved with conventional optimization methods. These problems can range from water resource optimization, control and optimization of renewable energy systems, supply chain management and risk management in sustainability proactive strategies. The quest for a solution to these kinds of problems has led researches to use soft computing techniques, by which they can obtain near optimal solutions [13-16]. Memetic algorithms are a class of meta-heuristic algorithms which combine a global optimization technique with a cultural evolutionary stage responsible for local refinements [5]. In this paper, we propose a novel adaptive memetic algorithm, where the influence factor of the environment, on

the learning abilities of each individual, is set adaptively. This translates into a level of autonomous behavior. In summary, we first propose a memetic algorithm, which includes an adaptive imitation scheme, and then we study the performance of variations of hill climbing search strategy in conjunction with the adaptive imitation. With the improved performance achieved by proposed memetic algorithm we can solve the above mentioned problems more efficiently and achieve better results.

3 Research Contributions and Simulation Results

As mentioned earlier, the structure of local search algorithms imposes a high computational cost, which is exponential. Simulation results verify that the performance of memetic algorithms is highly dependent to the neighborhood size. For example in a problem with continuous values for genes, smaller neighborhood radius (up to a threshold) leads to better solution, but the price is increased number of evaluations and much computational time. Hence, a dynamic imitation rate in which the neighborhood size is adjusted based on the iteration can guide the searching strategy more wisely. To calculate the imitation rate in continuous search space, equations below are presented:

$$\text{imitation_rate_I} = \left(1 - \frac{t-1}{T}\right)^a \quad (1)$$

$$\text{imitation_rate_II} = e^{-\frac{1}{2} \left(\frac{(t-1)^2}{aT} \right)} \quad (2)$$

where t is the current iteration, T is the maximum number of iterations and a is the adaption coefficient. Imitation rate diagram for different values of a is illustrated in figure 3.

The metaphorical concept suggests that imitation can be feckless in some situations, i.e. impulsive imitation misleads the evolution process. Each person during his life span benefits from the interactions with the environment as means to improve his competency in the society, but the scale to which he is influenced from the environment usually decreases as he grows older. In the proposed imitation adaption scheme,

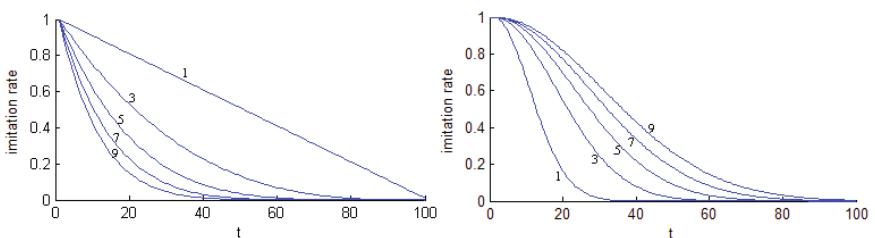


Fig. 3. Imitation rate for different values of a . Left: Equation 1, Right: Equation 2.

diversity of the population is estimated by the iteration number. In the first iterations population should be as diverse as possible to cover all searching space, but achieving final iterations diversity of solutions decreases. In this scheme the imitation rate of individuals decreases along the evolution process. This will not only reduce the unnecessary evaluation of neighbors, i.e. increase the speed, but it also regulates the exploitation rate and leads to improved solutions.

Hill climbing local search has been one of the most used canonical imitation operators, suitable for continuous search space [3,18,25]. In the following section different variations of hill climbing algorithm is evaluated and the simulation results are discussed. In these simulations the imitation rate is calculated adaptively using equation 1 and equation 2. To make comparisons we use Ackley test function. Numerous local minima in the search space of this function have made it one of the popular bench mark functions for the assessment of evolutionary algorithms [5,7]. Equation 3 represents the Ackley function in n -dimensional space:

$$f(X) = -ae^{-b\left(\sqrt{\frac{1}{n}\sum_{i=1}^n(x_i)^2}\right)} - e^{\frac{1}{n}\left(\sum_{i=1}^n \cos(cx_i)\right)} + a + e \quad (3)$$

$$a = 20, b = 0.2, c = 2\pi$$

Taking into account the stochastic nature of evolutionary algorithms, for each simulation we run the algorithm 30 times and get the best solution and average of the solutions to compare the algorithm performance. Variance of the solutions is also presented to assess the stability of algorithms; smaller variance indicates that solutions obtained from different runs of the algorithm are closer to the average index. Time index indicates the average time a single run of algorithm takes. For all memetic algorithms used in our study, properties of GA part are the same.

Complete hill climbing imitation-- In this algorithm all possible neighbors of a chromosome are generated and evaluated and the best neighbor replaces the current chromosome. The search continues until there is no any neighbor better than current state of the chromosome. The simulation results obtained from 30 runs of the algorithm are provided in table 1. To calculate the imitation rate equation 1 and equation 2 are used. Adaption coefficient is set empirically for each problem space. Since the number of iterations required for imitation operator is not predictable and the computational cost is exponential, running time for this algorithm differs in various runs and even the algorithm may end up acting like a complete search.

First-best hill climbing imitation-- Contrary to the complete hill climbing search, this algorithm stops the search procedure as soon as a neighbor better than current chromosome is found. In the worst case first-best hill climbing search is like complete hill climbing search, but the simulation results in table 2 show that the running time of this algorithms is improved compared to the previous one.

Table 1. Simulation results for complete hill climbing with adaptive imitation

Adaption equation	a	Time (s)	Solution average	Solution variance	Best solution
Eq. 1	3	16.02	4.368×10^{-8}	3.174×10^{-14}	0.000×10^{-3}
Eq. 2	5	16.02	8.839×10^{-8}	4.329×10^{-14}	3.553×10^{-15}

Table 2. Simulation results for first-best hill climbing with adaptive imitation

Adaption equation	a	Time (s)	Solution average	Solution variance	Best solution
Eq. 1	3	10.23	5.310×10^{-9}	6.672×10^{-14}	0.000×10^{-3}
Eq. 2	4	10.22	3.199×10^{-9}	5.648×10^{-14}	3.553×10^{-15}

Single-step hill climbing imitation-- This imitation operator considers a hill climbing search with only a single step (iteration). After all the neighbors of a chromosome are evaluated, best neighbor (if any) replaces the current chromosome and search stops. Since each imitation operation is a single step forward, the computational time of the algorithm in different runs is the similar. Simulation results are provided in table 3.

Table 3. Simulation results for single-step hill climbing with adaptive imitation

Adaption equation	a	Time (s)	Solution average	Solution variance	Best solution
Eq. 1	5	6.43	3.165×10^{-9}	4.677×10^{-17}	3.553×10^{-15}
Eq. 2	2	6.43	1.701×10^{-11}	3.032×10^{-22}	0.000×10^{-3}

Table 4 provides the simulation results to compare an enhanced GA with elitism, standard GA based memetic algorithm (MGA) and single-step memetic algorithm (SSMGA) together. It can be seen that the memetic algorithm with single-step hill climbing search and using the adaptive imitation outperforms other algorithms both in computational time and in the performance.

Table 4. Simulation results comparing GA and variations of MA with adaptive imitation

Algorithm	a	Time (s)	Solution average	Solution variance	Best solution
GA	-	4.32	4.09×10^{-6}	2.70×10^{-10}	5.58×10^{-15}
MGA	10^{-1}	11.2	1.10×10^{-6}	3.61×10^{-11}	5.55×10^{-15}
MGA	10^{-3}	285.2	7.49×10^{-7}	9.33×10^{-12}	3.55×10^{-15}
MGA	10^{-6}	3892	1.81×10^{-8}	1.67×10^{-13}	2.51×10^{-15}
MGA	10^{-9}	∞	-----	-----	-----
SSMGA	2	6.43	1.70×10^{-11}	3.02×10^{-22}	0.00×10^{-3}

4 Conclusions and Future Works

Memetic algorithms, being the most famous hybridization method, successfully improve the performance of their underlying global search algorithm. They effectively improve the stability and fine tune the algorithm to converge to the global optimum. Two major issues that MAs suffer from are increased computational time and designing appropriate imitation operator which defines the neighborhood function. Points that are locally optimal with respect to one neighborhood structure may not be with respect to another [12]. In this paper first we introduced an adaptive scheme to determine the imitation rate. In this scheme the capacity of individuals to be influenced by

the environment is reduced as they grow older. By studying different imitation structures, we introduce a single-step hill climbing search strategy coupled with adaptive imitation.

Improvements in quality of results and computation time are usually conflicting objectives. Simulation results show that the proposed approach not only improves the efficiency of the memetic algorithm, by fewer neighbor evaluations and less computing time, but it also improves the performance by achieving better results. This verifies that excessive imitation can potentially distract the algorithm from global convergence.

Proposed contributions have been developed for problems with continuous search space, using genetic algorithm and variation of hill climbing algorithm, but the concepts can be extended to discrete domain and using other metaheuristics. Future research will further address the automatic fine tuning of the adaption coefficient.

References

1. Krasnogor, N., Smith, J.: A memetic algorithm with self-adaptive local search: TSP as a case study. In: Proc. of GECCO, pp. 987–994. Morgan Kaufmann, San Francisco (2000)
2. Ong, Y., Keane, A.: Meta-Lamarckian Learning in memetic algorithms. IEEE Trans. Evol. Comput. 8(2), 99–110 (2004)
3. Ong, Y.S., Lim, M.H., Zhu, N., Wong, K.W.: Classification of Adaptive Memetic Algorithms: A Comparative Study. IEEE Transactions On Systems, Man and Cybernetics - Part B 36(1), 141–152 (2006)
4. Dawkins, R.: The Selfish Gene. Oxford University Press, New York (1976)
5. Eiben, A.E., Smith, J.E.: Introduction to evolutionary computing. Springer, Heidelberg (2003)
6. Engelbrecht, A.P.: Fundamentals of computational swarm intelligence. John Wiley, New York (2005)
7. Guimares, F.G., Campelo, F., Igarashi, H., Lowther, D.A., Ramrez, J.A.: Optimization of Cost Functions Using Evolutionary Algorithms With Local Learning and Local Search, Magnetics. IEEE Transactions (2007)
8. Freisleben, B., Merz, P.: New Genetic Local Search Operators for the Traveling Salesman Problem. In: Ebeling, W., Rechenberg, I., Voigt, H.-M., Schwefel, H.-P. (eds.) PPSN 1996. LNCS, vol. 1141, pp. 890–900. Springer, Heidelberg (1996)
9. Mariano, A., Moscato, P., Norman, M.: Arbitrarily large planar etsp instances with known optimal tours. Pesquisa Operacional (1995)
10. Araujo, M., et al.: Constrained Optimization Based on Quadratic Approximations in Genetic Algorithms. In: Mezura-Montes, E. (ed.) Constraint-Handling in Evolutionary Optimization. SCI, vol. 198, pp. 193–217. Springer, Heidelberg (2009)
11. Gallardo, J.E., Cotta, C., Fernández, A.J.: On the Hybridization of Memetic Algorithms with Branch-and-Bound Techniques. IEEE Transactions On System, Man, and Cybernetics, Part B: Cybernetics 37(1), 77–83 (2007)
12. Smith, J.E.: Coevolving Memetic Algorithms: A Review and Progress Report. IEEE Transactions on Systems Man and Cybernetics 37(1), 6–17 (2007)
13. Arabi, M., Govindaraju, R.S., Hantush, M.M.: Cost-effective allocation of watershed management practices using a genetic algorithm. Water Resource Research 42 (2006)
14. Donslund, G.B., Østergaard, P.A.: Energy system analysis of utilizing hydrogen as an energy carrier for wind power in the transportation sector in Western Denmark. Utilities Policy 16, 99–106 (2008)

15. McKinney, D.C., Lin, M.D.: Genetic algorithm solution of groundwater management models. *Water Resource Research* 30(6), 1897–1906 (1994)
16. Reca, J., Martinez, J.: Genetic algorithms for the design of looped irrigation water distribution networks. *Water Resources Research* 42 (2006)
17. Ong, Y.S., Lim, M.H., Chen, X.: Memetic Computation—Past, Present & Future [Research Frontier]. *IEEE Compu. Intel. Mag.* 5(2), 24–31 (2010)
18. Liu, B., Wang, L., Jin, Y.H.: An effective PSO-based memetic algorithm for flow shop scheduling. *IEEE Trans. Syst., Man, Cybern. B* 37(1), 18–27 (2007)
19. Krasnogor, N., Blackburne, B.P., Burke, E.K., Hirst, J.D.: Multimeme algorithms for the structure prediction. In: Guervós, J.J.M., Adamidis, P.A., Beyer, H.-G., Fernández-Villacañas, J.-L., Schwefel, H.-P. (eds.) *PPSN 2002. LNCS*, vol. 2439, pp. 769–778. Springer, Heidelberg (2002)
20. Krasnogor, N., Gustafson, S.: Toward truly ‘memetic’ memetic algorithms: discussion and proof of concepts. In: *Advances in Nature-Inspired Computation: The PPSN VII Workshops*, pp. 21–22 (2002)
21. Ullah, A.B., Sarker, R., Cornforth, D., Lokan, C.: An agent-based memetic algorithm (AMA) for solving constrained optimization problems. In: *IEEE Congress on Evolutionary Computation*, pp. 999–1006 (2007)
22. Ishibuchi, H., Yoshida, T., Murata, T.: Balance between genetic search and local search in memetic algorithms for multiobjective permutation flow shop scheduling. *IEEE Trans. Evol. Comput.* 7, 204–223 (2003)
23. Moscato, P.: On evolution, search, optimization, genetic algorithms and martial arts: toward memetic algorithms. In: Tech. Rep. Caltech Concurrent Computation Program, Rep. 826, California Inst. Technol., Pasadena, CA (1989)
24. Hart, W. E.: Adaptive Global Optimization With Local Search. In: Ph.D. dissertation, Univ. California, San Diego, CA (1994)
25. Bersini, H., Renders, B.: Hybridizing genetic algorithms with hill-climbing methods for global optimization: two possible ways. In: Proc. IEEE Int. Symp. Evolutionary Computation, Orlando, FL, pp. 312–317 (1994)
26. Ayanzadeh, R., Shahamatnia, E., Setayeshi, S.: Determining Optimum Queue Length in Computer Networks by Using Memetic Algorithms. *Journal of Applied Sciences* 9(15), 2847–2851 (2009)

Design and Applications of Intelligent Systems in Identifying Future Occurrence of Tuberculosis Infection in Population at Risk

Adel Ardalan¹, Ebru Selin Selen², Hesam Dashti³, Adel Talaat⁴,
and Amir Assadi³

¹ Department of Electrical and Computer Engineering,
University of Wisconsin, Madison WI 53706 USA

² Department of Applied Statistics and Applied Mathematics,
Izmir University of Economics, 35330 Balcova, TR

³ Department of Mathematics, University of Wisconsin,
Madison, WI 53706 Madison, WI 53706

⁴ Department of Animal Health and Biomedical Sciences,
University of Wisconsin, Madison WI 53706 USA

Abstract. Tuberculosis is a treatable but severe disease caused by Mycobacterium tuberculosis (Mtb). Recent statistics by international health organizations estimate the Mtb exposure to have reached over two billion individuals. Delay in disease diagnosis could be fatal, especially to the population at risk, such as individuals with compromised immune systems. Intelligent decision systems (IDS) provide a promising tool to expedite discovery of biomarkers, and to boost their impact on earlier prediction of the likelihood of the disease onset. A novel IDS (iTB) is designed that integrates results from molecular medicine and systems biology of Mtb infection to estimate model parameters for prediction of the dynamics of the gene networks in Mtb-infected laboratory animals. The mouse model identifies a number of genes whose expressions could be significantly altered during the TB activation. Among them, a much smaller number of *the most informative* genes for prediction of the onset of TB are selected using a modified version of Empirical Risk Minimization as in Vapnik's statistical learning theory. A hybrid intelligent system is designed to take as input the mRNA abundance at a near genome-size from the individual-to-be-tested, measured 3-4 times. The algorithms determine if that individual is at risk of the onset of the disease based on our current analysis of mRNA data, and to predict the values of the biomarkers for a future period (of up to 60 days for mice; this may differ for humans). An early warning sign allows conducting gene expression analysis during the activation which aims to find key genes that are expressed. With rapid advances in low-cost genome-based diagnosis, this IDS architecture provides a promising platform to advance Personalized Health Care based on sequencing the genome and microarray analysis of samples obtained from individuals at risk. The novelty of the design of iTB lies in the integration of the IDS design principles and the solution of the biological problems hand-in-hand, so as to provide an AI framework for biologically better-targeted personalized prevention/treatment for the high-risk groups. The iTB design applies in more generality, and provides the potential for extension of our AI-approach to personalized-medicine to prevent other public health pandemics.

Keywords: Mycobacterium tuberculosis, biomarkers and intelligent decision systems, early detection, Vapnik's statistical learning theory.

1 Introduction

Tuberculosis is a severe lung disease which is responsible for increase of 9.4 million [1] cases per year that will result about 2 millions of patients death [2]. Infection will be acquired by inhalation of Mycobacterium tuberculosis contaminated air and/or droplets [3]. Even from slight initial invasion of the agent, infection may lead to latent TB or may lead to primary disease [4]. The exact time which is taken from the initial infection until the development of disease varies among individuals. This variation can be attributed mainly to the immune status of an individual. Immuno-compromised individuals or individuals which their immune systems are suppressed might likely develop primary disease. The World Health Organization WHO [1], has reported the burden of disease under various circumstances, and in particular, the number of bacteria sufficient to infect an individual [3]. After 2008, the numbers of incidence, prevalence and mortality, are estimated to be at least as 9.4 million incidence cases, 11.1 million prevalence cases and 1.3 million deaths [1]. Since an individual can infect 10-15 individuals [3], from a public health viewpoint, early diagnosis and treatment [5] is crucial to contain the disease. In particular, technologies for early detection and isolation of an infected individual will play a major role in sustainability of the global population health.

Understanding the molecular mechanisms during the invasion of *M. tuberculosis* provides valuable insights for the analysis of the biological understanding of the course of infection. Besides, the molecular understanding might lead to development of the necessarily better targeted treatment strategies against tuberculosis [6]. To serve this purpose, Talaat et al. (2004) analyzed MTB infected lung samples of immuno-compromised and immune-competent mice. They found the genes that are expressed (or significantly changed) during early invasion through systematic application of the microarray technology [7]. As a result of this analysis, they reported the differences between expression profiles in three different environments. In an analogous fashion, but in a different context, microarray technology is used to monitor the changes in *M. tuberculosis* gene expression during the treatment with antituberculous drug isoniazid [8]. According to analysis of expressed genes in presence of isoniazid, researchers are more likely to enhance the drug targets. Behr et al. (1999) perform microarray analysis in order to understand the differences between genomic structures of *M. tuberculosis* and *M. bovis* and other strains of *M. bovis* which are also compound of the BCG vaccines. Accordingly, this study has aimed to serve the purpose of developing new and more narrowly-aimed vaccines and/or antituberculosis treatment [9]. Fisher et al. (2002), draw attention to the function of the acidification during the immune response through using microarray [10]. As a result of their analysis, they suggest that, acidification might be a signal to induce the gene expression needed by the bacteria to survive against the immune response cells known as phagosomes.

The discussion above is short, but essentially highlights the critically sparse state-of-art knowledge regarding detection, prediction and treatment of individuals at risk, and in fact, almost all categories of individuals infected by this microbe.

In the following, we report on preliminary progress in design of an intelligent system (Section 3) based on our earlier *de novo* analysis of gene expression time-series by novel applications of stochastic signal processing, new clustering algorithms, and dynamics of representations of clusters in an appropriate hyperbolic space.

2 Contribution to Sustainability

One remark encountered often in modern higher education and research is the significance of close collaboration between young investigators in the computational and the engineering sciences with biologists to provide a fruitful framework for the synthesis of diverse concepts and tools. In this way, integration of hardware-software and ‘biological knowledgeware’ provides prospects of imminent solutions of myriad challenging biological sustainability research problems through collaborative effort. This research is an illustration of this piece of educational wisdom in this regard. The need for solution of hard biological problems has inspired formulation of a number of challenging problems in scientific computation and engineering.

High performance scientific computation plays a critical role in the 21st century research and engineering design optimization. An important case arises in research on challenging problems in Global Public Health, such as sustainability of protective measures against infectious diseases for humans and animals living across all geographic locations, and as much as possible, under all states of living conditions. This research adheres to the above-mentioned objective of sustainability, utilizing the state-of-art in informatics and engineering. Design of intelligent systems have enriched the modern technological societies, and extending its domain to include the entire global community is inevitable for future protection of life and earth and assurance of a sustainable mechanism to provide a healthy society across the globe. Further, among myriad health and disease conditions, there is a serious risk of faster spread of diseases such as TB by more susceptible sectors of individuals at risk due to a compromised immune systems. Sustainability of global health, therefore, must include effective solutions to prevent infection and more likely death by such individuals. On the ethical side, there are limitations to keeping in isolation individuals-at-risk due to the higher risks of eventual severe or fatal sickness. The research on iTB system provides a viable approach to answer the above-mentioned problems based on effective applications of modern engineering and informatics.

3 iTB: Intelligent TB Dynamical Modeling

We have developed a conceptual framework for dynamical analysis on the grounds of solid mathematical models and empowered by software engineering viewpoint toward information systems development. This viewpoint ensures reusability of the system as a whole or in part for different applications in different disciplines. In other words, a *modular* design of the system allows us to (1) simplify extending the system capability while maintaining accuracy, (2) rebuild new configurations on demand - e.g. for different clinical applications, and (3) distribution of massive computation among different processing blocks. The entire setting adheres to the engineering principles for deployment within today’s High Performance Computation (HPC) infrastructures,

like Computational Grids, and most recently, Computation in Clouds. Data-intensive computing era calls for such adaptable architectures to guarantee the applicability of informatics methods for growing population, increasing healthcare demands and personalized (thus, real-time) medical treatment.

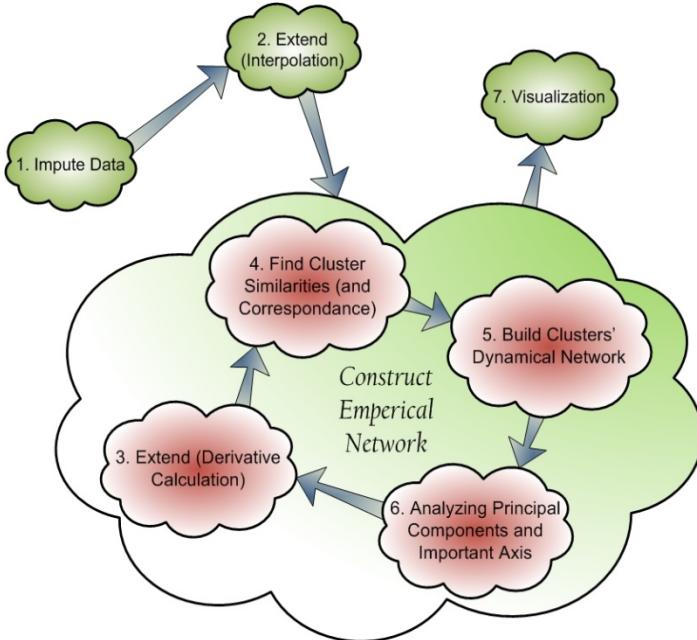


Fig. 1. Global architectural view of iTB

Fig. 1 shows a global architectural view of iTB. The main part is the *empirical network construction* module which can work iteratively to gain computational access into higher levels of dynamics of massively complex biological systems.

A brief description of modules is provided below. The *Imputation Module* tries to fill in the missing pieces of the data, which is a common issue in biological time-series samples. The algorithm utilized here relies on the widely-used Expectation Maximization (EM) algorithm to find the best candidates (i.e. most probable ones) for missing values. We have adapted a generalized version of EM [11] to apply to our problem setting. The next Module also performs preprocessing provides a suitable smoothing of the time-series for mathematical analysis, which requires stable local behavior of the time-series (regarding variations and disregarding spiky results). We have employed piecewise cubic Hermite interpolating polynomial (PCHIP) curve fitting approach [12] and re-sampled the regressed curve to approximately simulate the behavior of the system in an appropriately *smoother* way to accommodate differentiation and other analytic operations.

The next few Modules are designed to find the multi-level structure of the complex networks. They deal with the twisted behavior that results from the connectionism beyond the “*build from the simplest blocks*” philosophy. In non-technical terms,

connectionist and other learning-theoretic models are constrained by the nature of the domain of generalization, and the balance between the sufficiency of samples versus overtraining. This implies the requirement of elucidating the interrelationship among samples of different levels of significance for estimating future dynamics from the sample point (among other technical hurdles to overcome the biological complexity in predicting the state of the disease from a sparse sample.) Thus, to capture the dynamics of the disease, hierarchical clustering techniques are employed to build a multi-level structural/behavioral model of interactions inside the system. There are advantages to using hierarchical clustering versus non-hierarchical clustering. A comprehensive analysis of different clustering methods and their applicability will appear in due time. Briefly, hierarchical clustering allows the modeler to take into account various forms of analytic singularities, avoiding the artificial assumption that the data samples are uniformly chosen from a single probability distribution function (pdf). Hierarchical clustering, on the other hand, offers the more realistic assumption that different sub-clusters are samples from several pdf that could somewhat differ from the initial pdf. Such diversity of pdf is expected in biology, due to variation and other complex biological phenomena.

The remaining Modules attempt to render the novel concepts of *special-architecture* empirical networks for topological modeling of complex networks. This approach enables us to use a vast spectrum of solid mathematical analysis tools to reveal invaluable measures of correspondence between components of complicated systems (cf. below for an outline.)

Two of the Modules capture dynamic similarities, and record the migration of different classes of genes with different perspectives towards the inclination of the groups' rate and acceleration change in the course of time. The next Module estimates the probability densities in clusters via the exponential family of models. The exponential family has the unique advantage of being quite flexible to accommodate many deviations from Gaussian models, while still are parameterized via a finite dimensional Riemannian manifold in the Hilbert space of L^2 -integrable functions. The Riemannian structure alluded above is complete and hyperbolic. Hypothesizing the consistent behavior of related genes in the hyperbolic space, we have measured the distance between the clusters of genes as follows, which we mention in the Gaussian case to simplify the presentation.

The individual clusters are regarded as samples from a probability distribution; e.g. for the sake of a concrete illustration, consider two clusters that are regarded as samples from two pdfs that are normal distributions

$$\mu_j \sim N(m_j, \sigma_j^2) \quad (1)$$

So for any time frame, we have a topological representation of the system that evolves in time and shows the behavior of the systems in the frame of the hypothesis. To be able to visually inspect the dynamics of the system, we may argue based on the experimental results that projection of the high-dimensional networks obtained into the very first principal components gives us a good representation of the behavioral model.

3.1 Mathematical Methods

First, the solution of the preceding estimate requires *transforming the ill-posed problem* into a regularized well-posed problem [13], [14] and [15]. Thus, it is desired here to have a well-posed problem regarding *estimating the measure* on the “function-space”. There are well-known methods and more modern ongoing research on regularization, and we shall omit the discussion due to lack of space [13] and [15]. To readers familiar with learning theory [16], the latter problem could be regarded as “*Learning the Measure*” from the sample of trajectory dynamics (we used the data available to us from the TB-infection of mice) through a controlled iterative scheme. Thus, we proceed to cast the latter in *Statistical Learning Theory*. Accordingly, we need to have a robust estimate for the error in iterative steps of learning to quantify the approximation error for the posterior mentioned above. Robust error estimates require stability in solution of sampling. To have a well-posedness of the inverse problem for the posterior measure will provide desired levels of stability. In turn, such stability may be used as the basis for quantifying the approximation of inverse problems for functions in a finite-dimensional-space setting. This requires an estimate for the distance between the true and approximate posterior distributions, in terms of error estimates for approximation of the underlying forward problem.

Let μ_1 and μ_2 be two normal distributions with means m_1, m_2 and standard deviations σ_1 and σ_2 . The computation of distance between them $d_H(\mu_1, \mu_2)$ is as follows

$$\left(1 - \sqrt{\frac{2\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2}} e^{-\frac{1}{4} \frac{(m_1 - m_2)^2}{\sigma_1^2 + \sigma_2^2}} \right)^{\frac{1}{2}} \quad (2)$$

This metric is initially defined for two measures μ_1 and μ_2 that are absolutely continuous relative to a measure λ as:

$$d_H(\mu_1, \mu_2) = \left(\frac{1}{2} \int \left(\sqrt{\frac{d\mu_1}{d\lambda}} - \sqrt{\frac{d\mu_2}{d\lambda}} \right)^2 d\lambda \right)^{\frac{1}{2}} \quad (3)$$

where $\frac{d\mu}{d\lambda}$ terms are Radon-Nikodym derivatives respectively. The definition of (3) does not depend on the choice of λ , so (3) does not change if λ is replaced with a different probability measure with respect to which both μ_j are absolutely continuous. For two normal distributions (1), the formula for (3) is (2) and readily could be used, even helped by symbolic algebra, to improve accuracy of the iterative Learning-theoretic calculations. Control of differences in the Hellinger metric (i.e. d^2_H) leads to control on the differences between expected values of functions and operators (that admit polynomial bounds). Statistical Learning Theory [16] provides the tools to complete the remaining steps in this approach. There are other computational reasons for choice of the Hellinger metric versus other probability-theoretic divergences, say

from the family of f-divergences such as the Kullback-Leibler, Wasserstein or other metrics. Among them, one could gain useful estimates of bounds more easily using this metric, such as in

$$\left\| \int f d\mu_1 - \int f d\mu_2 \right\| \leq 2 \left(\int \|f\|^2 d\mu_1 + \int \|f\|^2 d\mu_2 \right)^{\frac{1}{2}} d_H(\mu_1, \mu_2) \quad (5)$$

In turn, such bounds point to design of numerical schemes that allow us to solve the inverse problem and gain control in relating estimates arising during perturbations in the domain and range, respectively. Briefly, in the discussion above, let the integer N denote the iteration count, and correspondingly, the estimates $\Psi^N(v;w)$, μ^N , such that $\frac{d\mu}{d\mu_0} \sim e^{-\Psi(v;w)}$, $\frac{d\mu^N}{d\mu_0} \sim e^{-\Psi^N(v;w)}$. Then bounds on (6) provides a sequence of improving bounds on (3) that demonstrates when $N \rightarrow \infty$, the Hellinger metric approaches zero exponentially in N .

$$|\Psi(v;w) - \Psi^N(v;w)| \quad (6)$$

For Gaussian densities, these bounds are used to prove that the means and covariance operators associated to μ^N , and μ , are close in the Hilbert-space (or the Banach space, in more general circumstances) operator-norms. Therefore, in the approach outlined above, we could arrange for the transfer of estimates from the numerical analysis of forward problems into estimates for the solution of the related inverse problem.

4 Results and Discussion

In the preceding arguments, the space of exponential probability density functions is a Riemannian manifold, and we need a discrete approximation to capture its metric properties within the prescribed error bound. The discrete approximation is typically expected to be high dimensional (thousands or more). In the case of normal distributions, a number of analytic simplifications are available that allow of us to reduce the dimensionality of the metric model. In particular, the Riemannian metric could be approximated by sampling of the distance data, and the model reduction for the sample agrees with the desired approximation to the Riemannian structure. As expected Singular Value Decomposition provides the direct approach, and in the case of animal models of the disease (murine), the results are obtained as follows. Empirically, we have observed that considering the first three principal components retains about 80-85% of the information content of the network (i.e. the ratio of the first 3 eigenvalues to the total sum), while the dynamical separation of different conditions under study are brilliantly visible. Figures 2 to 7 show a sample dynamics of the TB genes in the level of data. The axes are all relative to a unit-free representation of the cluster distances as measured in the Heilinger distance. Figures are converted to the 3-dimensional projection metric for visualization purposes. The figures are shown in different zoom levels for clarity purposes.

We have studied, as an example, the dynamics of gene expression profiles of TB-related families in a period of 24 hours after infection. Profiles are recorded every 4 hours and a sliding window approach is used to investigate the differences between

the “behavioral associations” among genes. As could be observed in the figures, there is a heart-beat-like pattern between clusters representing the conditions. Each point on the graphs represents a cluster of genes in one of the conditions. The observable dynamics resemble a group of particles in a force field which approach each other and then the repulsing forces cause the system to scatter around.

The importance of the above-mentioned results is that the behavioral differences between the two conditions are observable in the very first stages of getting exposed to the invader. The discrimination between the two groups of points (red crosses vs. blue circles) which correspond to the different conditions, could be observed from these graphs as a growth/shrinkage pattern. From a clinical treatment viewpoint, this is of utmost importance as one could be treated before the infection spreads out of control.

At the time of writing this article, we are making progress in translating the above-mentioned visualization-based observations (the differences between the two cases) into a classification scheme within the 3-dimensional principal component space (please see figures). Our method is based on design of new algorithms that minimizes the empirical risk function (analogous to the Vapnik soft-margin SVM empirical risk functional), using the hyperbolic metric in lieu of the Euclidean distance in the theory by Vapnik and others. While the computations are much more challenging, the mathematical results that must guarantee the existence of a minimum for the risk functional and the desired regularity properties are ensured through extensive mathematical work on analysis of functions on hyperbolic spaces, in particular, bounded sequences of approximations to a minimum converge exponentially (hence even faster than the Euclidean metric) and there is a unique limit point for the sequence, hence a unique minimum. The details will be provided in a forthcoming paper.

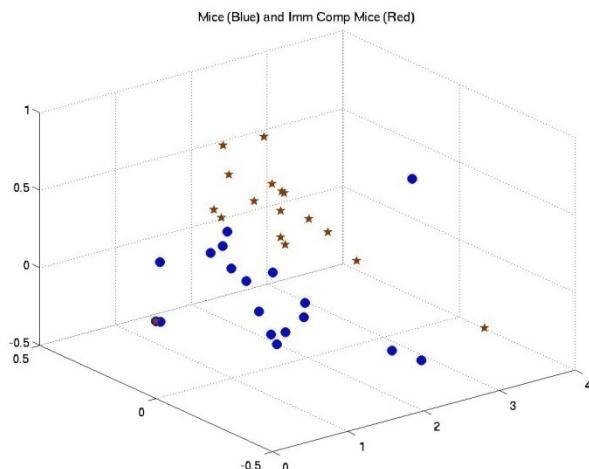


Fig. 2. These samples of the visualization movie frames show instances of patterns in dynamics of the hyperbolic representation of the analyzed sample of TB gene expressions in time. The dynamics is approximated discretely, then projected to the 3-dimensional reduced model from the hyperbolic space. The first three dominant principal modes capture more than 80% of the information contents in the original hyperbolic space. Once the separations of different dynamic patterns are accomplished in the reduced model, clearly the original data will also demonstrate the separation by considering the inverse-images of the separating hyperplanes.

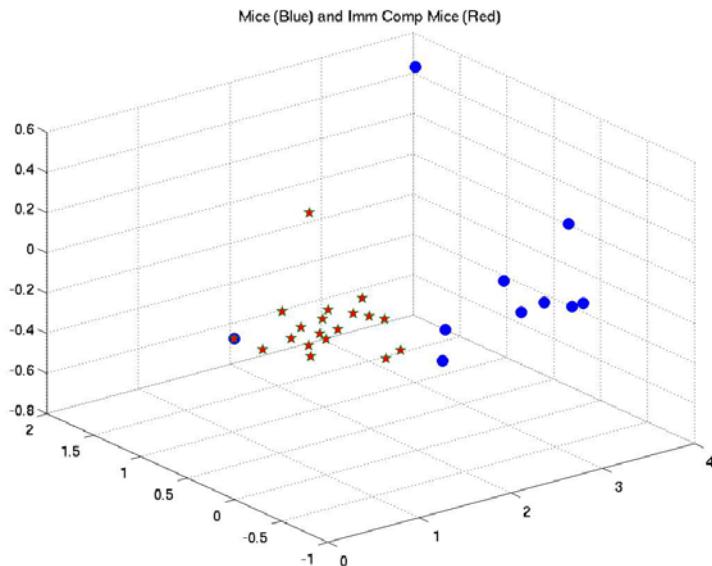


Fig. 3. A sample of projection of the gene expression dynamic pattern in the reduced model

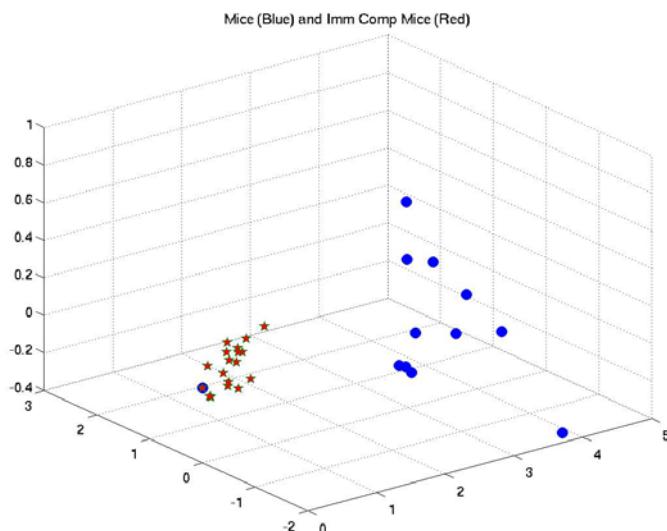


Fig. 4. A sample of projection of the gene expression dynamic pattern in the reduced model

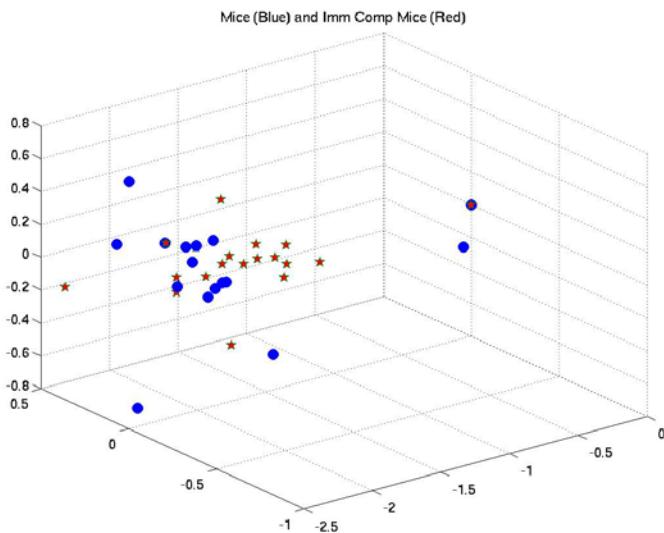


Fig. 5. Another sample of projection of the gene expression dynamic pattern in the reduced model

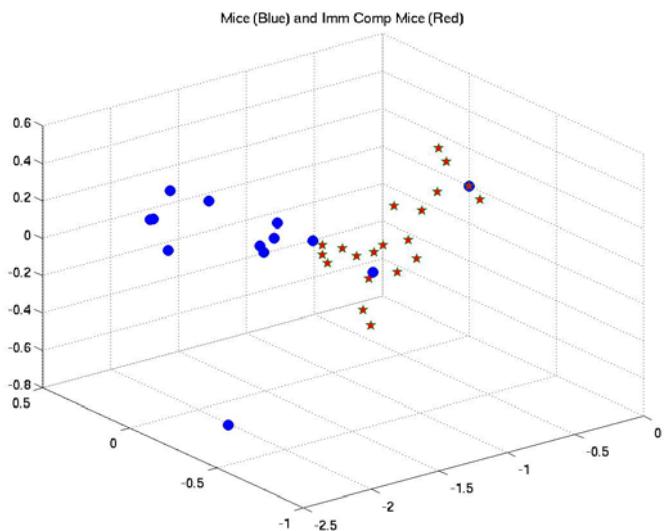


Fig. 6. Another sample of projection of the gene expression dynamic pattern in the reduced model

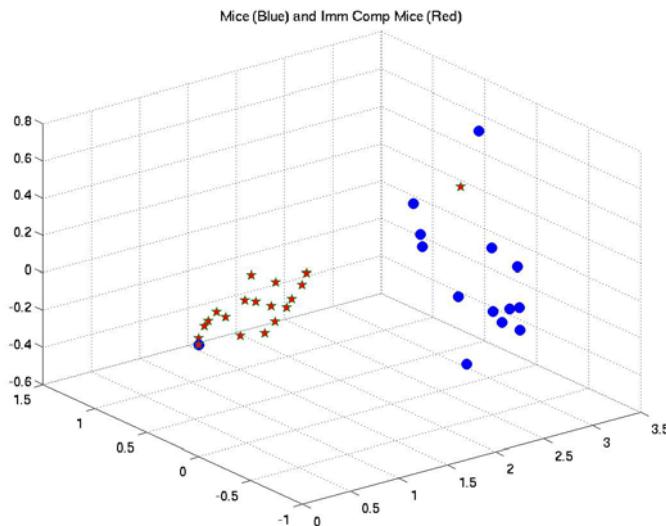


Fig. 7. Another sample of projection of the gene expression dynamic pattern in the reduced model

References

1. World Health Organization: Global Tuberculosis Control: A short update to the, report (2009)
2. Volokhov, D.V., Chizhikov, V.E., Denkin, S., Zhang, Y.: Mycobacteria Protocols. Humana Press, New York (2008)
3. World Health Organization, <http://www.who.int/topics/tuberculosis/en/>
4. Murray, M.: Tuberculosis: The Essentials. Informa Healthcare (2010)
5. Hopewell, P.C.: Tuberculosis: The Essentials. Informa Healthcare (2010)
6. Triccas, J.A., Berthet, F.X., Pelicic, V., Gicquel, B.: Use of Fluorescence Induction and Sucrose Counter Selection to Identify *Mycobacterium tuberculosis* Genes Expressed Within Host Cells. *Microbiology* 145, 2923–2930 (1999)
7. Talaat, A.M., Lyons, R., Howard, S.T., Johnston, S.A.: The Temporal Expression Profile of *Mycobacterium tuberculosis* Infection in Mice. *Proc. Natl. Acad. Sci.*, 4602–4607 (2004)
8. Wilson, M., DeRisi, J., Kristensen, H.H., Imboden, P., Rane, S., Brown, P.O., Schoolnik, G.K.: Exploring Drug-induced Alterations in Gene Expression in *Mycobacterium tuberculosis* by Microarray Hybridization. *Proc. Natl. Acad. Sci.*, 12833–12838 (1999)
9. Behr, M.A., Wilson, M.A., Gill, W.P., Salamon, H., Schoolnik, G.K., Rane, S., Small, P.M.: Comparative Genomics of BCG Vaccines by Whole-Genome DNA Microarray. *Science* 284, 1520–1523 (1999)
10. Fisher, M.A., Plikaytis, B.B., Shinnick, T.M.: Microarray Analysis of the *Mycobacterium tuberculosis* Transcriptional Response to the Acidic Conditions Found in Phagosomes. *J. Bacteriol.* 184, 4025–4032 (2002)
11. Schneider, X.: Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. *Journal of Climate* 14, 853–871 (2001)

12. Fritsch, F.N., Carlson, R.E.: Monotone Piecewise Cubic Interpolation. *SIAM J. Numerical Analysis* 17, 238–246 (1980)
13. Poggio, T., Girosi, F.: Regularization algorithms for learning that are equivalent to multi-layer networks. *Science* 247, 978–982 (1990)
14. Girosi, F.: An Equivalence Between Sparse Approximation and Support Vector Machines. *Neural Computation* 10, 1455–1480 (1998)
15. Smola, A.J., Schölkopf, B.: Form Regularization Operators to Support Vector Kernels. Morgan Kaufmann, San Francisco (1998)
16. Vapnik, V.N.: *The Nature of Statistical Learning Theory*. Springer, New York (2000)

Gait Intention Analysis for Controlling Virtual Reality Walking Platforms

Laura Madalina Dascalu¹, Adrian Stavar¹, and Doru Talaba²

^{1,2} Transilvania University of Brasov, Product Design and Robotics Department,
Bulevardul Eroilor, nr. 29, 500036, Brasov, Romania
{madalina.dascalu,adrian.stavar,talaba}@unitbv.ro

Abstract. Simple human gait can be difficult, unfeasible and not always practical in Virtual Reality, because of spatial and technological limitations. For 3D virtual environment travelling, different walking platforms have been developed in the last few years, but navigation using most of them is far from bringing naturalness to the user's movements. Users sometimes are unsecure and they are trying to adapt and to correct any irregularities they feel. Our research is focused on specific walking patterns that characterize the intention of walking: starting walking with a certain speed, maintaining a desired speed, accelerated walking, decelerated walking, stopping. In laboratory conditions, using a motion capturing system, these behaviors were reproduced, measured and analyzed.

Keywords: human gait, Virtual Reality, walking patterns, walking platforms, walking intention, gait analysis, kinematics.

1 Introduction

The main motor activity of daily life, human walk, is a complex, dynamic and unforeseeable process. It involves multiple joints movements, demands sensory-motor integration and the synchronization of the skeleton with the neurological system. Even if it runs mostly unconsciously, it has important goals: to keep upright, to maintain balance, to avoid collapse, to move the center of mass forward, change direction when necessary, avoid obstacles, adapt for avoiding painful forces or motions etc [1]. Human gait is integrated in a big, autonomous and auto-adaptive control system, capable of learning, which is the human being.

The complexity and difficulty of human gait have made it difficult to use as a key element present in the natural human-machine interaction. In Virtual Reality (VR), human gait is considered the user's most natural way of exploring a virtual environment [2], [3]. The user does not have to learn any interaction metaphor [4]; he can use his own walk for navigating inside the virtual environment, in the same way as he does in the real environment.

For Virtual Reality travelling, multiple devices and platforms have been built, to facilitate the use of human natural walk as the main way of navigating and exploring a 3D virtual environment. Though, most existing platforms are either too large or

expensive, or don't bring naturalness and safety to the users' movements during travelling [5], [6], [7], [8]. To create to the user the perception of natural walking in a virtual environment, a VR travelling platform should permit the user to walk freely, in any direction and with any desired speed, but this is not always possible because of physical and technical limitations.

The control system must have an appropriate response regarding the user's locomotion, in order to assist him in a freely and unrestricted manner, but at the same time to keep him/her as precise as possible in a fixed point and area of the walking surface. In this case one of challenges for the controlling system is related to the various changes of speed and direction in human natural walking. The main question is whether it is possible or not to anticipate and predict the gait intent in order to allow a real time control for high precision gait cancelling.

The attention of this paper is focused on the various changes of the linear walking speed. It is presented an approach for gait intent identification based on classification of the human postures while taking various decisions for linear walking. Analyzing the specific walking inputs that the control systems of travelling platforms use [5], [6], [7], [8] we observed that they are not focused on specific decision patterns that characterize the intention of walking: starting walking with a certain speed, maintaining a desired speed, accelerated walking, decelerated walking, stopping. These locomotion intentions can be externally observed in user's walking behaviors. In laboratory conditions, using a motion capturing system, these decision walking behaviors were reproduced, measured and analyzed.

2 Contribution to Sustainability

Virtual Reality is one of the most promising simulation technologies for the future sustainability. The application areas range from medical, engineering to urban planning, training and education, sports and arts. Natural navigation of the human user in Virtual Environments is one of the current bottlenecks which are expected to unlock a large number of applications on the entire spectrum but especially in the area of urban planning and built environment which are crucial for the future sustainable development.

One of the latest applications of gait study concerns the natural navigation in virtual environments using special devices like carpets and treadmills that are able to cancel the human displacement while walking. The main problem for the control of these devices is the identification of human gait intent in order to compensate in real time the displacement with the highest possible precision. Usually, the latency in anticipation of the displacement intent leads to the necessity of a device with a larger walking area. The more precise is the anticipation of the gait intent the smaller the needed walking area is needed. In case of Virtual Reality applications with limited user space, increasing the speed and precision of anticipation is crucial, as the user have a limited physical space available for walking.

The proposed approach, presented in this paper, is focused on building better controlling systems for Virtual Reality travelling platforms that will adapt to the user needs and will eliminate the user's intervention for machine accommodation. This paper presents results of preliminary tests for linear walking, which are promising for

further development of our project: obtaining the instantaneous parameters for omnidirectional walking. Nevertheless, identifying the gait intent is a very complex issue that not only needs much more information about the instantaneous postures, but also interpretation of these data such as to be able to predict the next moves.

3 Methodology

3.1 Participants

Five healthy young subjects were recruited as volunteers from the student population of Transilvania University of Brasov, Romania. Before participating in the experiments, the subjects were fully informed about the nature of the study and they all gave their consent to participate to it. The mean age of the participants was 25. The participants wore comfortable shoes and tight fitting clothing for reducing the movements of the markers placed on the body (Fig. 1).

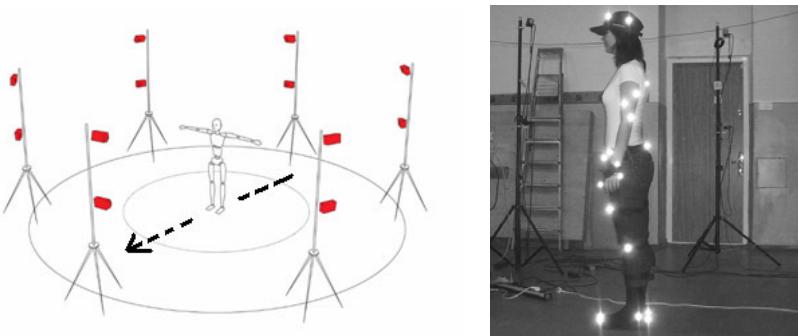


Fig. 1. Left: Camera placement (*IR cameras in portrait orientation*), capturing area and subject's walking direction (dotted line). **Right:** Full body marker placement (*lighting points*).

3.2 Apparatus

Human walking was measured using the Optitrack Motion Capturing System [9]. Optitrack Motion Capturing System is a passive optical system that use markers coated with a retro-reflective material to reflect light back to the infrared cameras (IR). Twelve IR cameras were used for data capture (as it can be seen in the Fig. 1).

The optical motion capturing system was chosen for the main advantages it has: high update rates, low latency and scalable to fairly large areas [10]. The surrounding environment was designed carefully to reduce ambient radiation, the main disadvantage of this motion capturing system.

A stationary laboratory coordinate system was defined by a vertically oriented Y-axis, a Z-axis placed forward in the walking direction, and an X-axis perpendicular to the first two. A calibration volume was set to 2.50, 2.50, and 2.20m. The motion analysis model used was a full body human model with 34 markers set, placed on specific landmarks [9], [10] of the body (full body marker placement is shown in

Fig. 1). For higher precision in marker capturing a skeleton was used. The skeleton tracking rate is of 100 FPS (frame per seconds) for the FLEX:V100 cameras used in the presented study.

3.3 Test Scenario

All testing were performed in the Virtual Reality Research Laboratory from Transilvania University of Brasov, Romania. Subjects had to perform some overground linear walking on the laboratory floor, according to specific received instructions. The instructions received by the participants refer to: starting and stopping linear walking motion at their self-selected moment of time, performing some linear walking at their desired speed and then accelerate (at their self-selected speed) or decelerate (at their selected speed), performing an accelerated linear walking and then decelerate, performing a decelerated linear walking and then accelerate.

The aim of this experimental research was to reproduce, in a controlled environment, the linear walking movements a user does when exploring an environment (it can be real or virtual). There were considered only a limited number of linear walking states, being selected only those considered relevant for VR navigation. These linear walking states with their characterization are presented in Table 1. There were also analyzed the transitions from one state to another, in the same context discussed above.

Table 1. Linear walking states

Linear walking state		State characterization
Repose	R	The subject is in repose state when his velocity $v = 0$.
Constant walking	CW	The subject is in constant walking state when the mean value of its Center of Mass (CoM) velocity is constant
Accelerated walking	AW	The subject is in accelerated walking state when the mean value of its Center of Mass velocity is increasing from one gait cycle to the next one
Decelerated walking	DW	The subject is in decelerated walking state when the mean value of its Center of Mass velocity is decreasing from one gait cycle to the next one

4 Results and Discussion

All experiments were conducted in controlled laboratory conditions. Specific walking scenarios were reproduced and measured using real-time motion capture data from a marker-based optical motion capture system, the Optitrack Motion Capturing System. Kinematic procedures are used for measuring the spatial motion of the full body and of its segments, during the representative linear walking states. Temporal and spatial kinematic gait parameters provide information about the time and position of the persons' gait.

Previous studies analyzed gait initiation and gait termination in terms of balance and control [11], [12], [13] but little was done in terms of gait initiation with the aim of controlling travelling platforms for VR. Gait initiation represents the transition from the repose state (see Table 1) to the CW state, AW state or DW state, all of them, being a periodic movement from one base of support to another, the gait cycle.

A gait initiation sequence can be observed in Fig. 2. Analyzing the lower extremities of the body, the first movement in gait initiation can be observed at the knee level, but the transition movement from the repose state to first heel contact (HC) of gait cycle can be observed only at foot level, when the right foot is up and the left foot is in inverted pendulum support, supporting the body weight. The left foot initial push is moving the body's CoM forward and to the right [11]. This is forcing the right leg to get from swinging to heel contact, characterized by the heel marker position close to the ground.

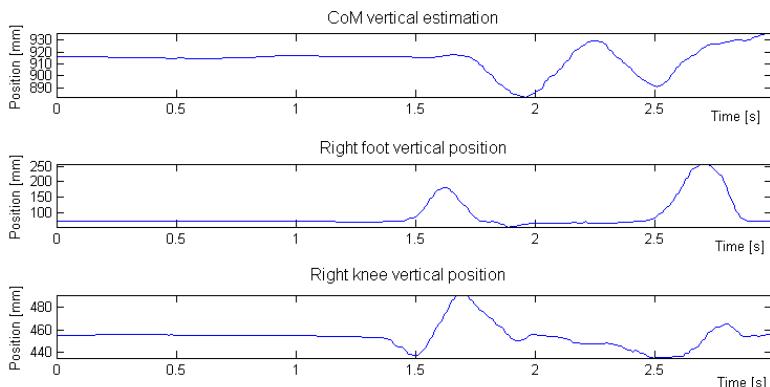


Fig. 2. Lower extremities marker displacements during gait initiation

The desired speed is achieved only after the next heel contact of the right leg. The periodicity of the same markers is confirming this statement. The autocorrelation function for the knee marker displacement shows a periodic movement, as it can be seen in Fig. 3.

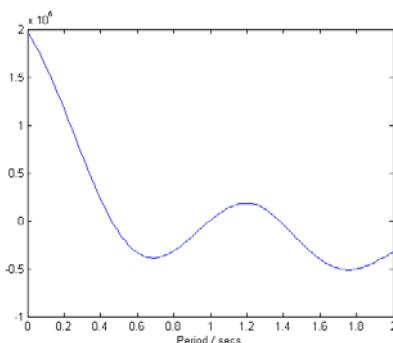


Fig. 3. Estimated periodicity of the movement of the knee marker

Autocorrelation function for the heel and knee markers revealed no periodic movement along both the vertical and horizontal dimensions at slow walking motions, but only on trunk marker, which can be theoretically explained by the periodicity of the CoM displacement during gait cycle [1], [14], [15], [16].

Gait termination (normally on double feet support) involves on one side, stance phase from the associated leg, completed on the other side with a shortened swing phase from the opposite leg and its placing next to the first one. The stopping sequence is much difficult to measure in terms of prediction using specific kinematic gait patterns. Coming back to the repose state is achieved only through a decelerated walking state which can be observed. The effective gait termination can be on the same foot with gait initiation or on the opposite foot. Fig. 4 shows a return to the repose state from the perspective of the stance support leg and of the opposite leg.

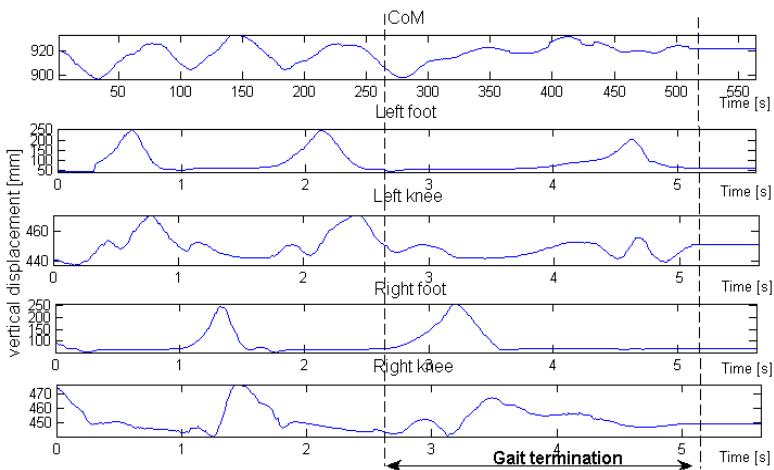


Fig. 4. Estimated periodicity of the movement of the knee marker

Gait termination presented in Fig. 4 starts with a heel contact of the left foot – the stance support leg. At this moment all the feet parameters are the same as in the periodic gait cycle, but it can be observed a smaller displacement at the level of the right knee. This leads to the idea of taking into account the position of the both feet when analyzing the transition from walking to repose state.

A left foot balance can be also observed after the heel contact of the opposite leg. This can be theoretically explained by applying a backwards and leftwards force to the body center of gravity, arresting its forward motion and bringing it to the midpoint between the feet [11]. Even it has been shown that there are asymmetries between the two sides of the body [17] for the purpose of this study these asymmetries can be neglected.

A transition sequence, from CW to AW, and then to DW and stopping can be observed in Fig. 5 in terms of velocities. The increasing of the shank's speed is come along with an increasing of the user's speed. Also, a higher decreasing of shank's speed is necessary for decreasing the user's speed.

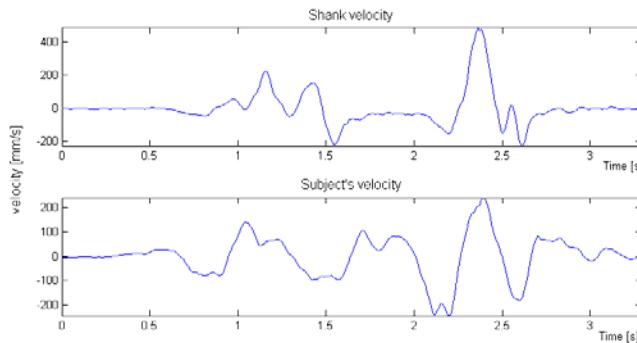


Fig. 5. Estimated periodicity of the movement of the knee marker

5 Conclusions and Future Work

The presented research from this paper represents the early results of an ongoing research project at Transilvania University of Brasov and brings into attention the identification and prediction of human linear walking for building better controlling systems for Virtual Reality walking platforms. The aim of the present study was to get a measure of the natural linear walking, in transitions from repose to steady-state continuous walking, or from accelerated walking to decelerated walking and then to repose again etc. There were of interest only the aspects of the human linear walking; the omni-directional walking parameters in the context discussed are subjects of our further research.

The method described in this paper and the proposed directions in identifying user intent aims to achieve to a new level of understanding human gait, dedicated to real-time prediction of gait intention and real-time control and command of VR platforms inside an immersive environment. This level is centered on observable intentions of human walking: the measured human motor actions.

Acknowledgments. This paper is supported by the Sectoral Operational Programme Human Resources Development (SOP HRD), financed from the European Social Fund and by the Romanian Government under the contract number POSDRU/6/1.5/S/6 for the authors (1) and by the research project IREAL contract no. 97/2007, id:132, funded by the Romanian Council for Research CNCSIS for the authors (2).

References

1. Vaughan, C.L., Connor, O., Davis, J.C., Dynamics, B.L.: *Human Gait*. PublishersSmith, Kiboho (1999)
2. Usoh, M., Arthur, K., Whitton, M.C., Bastos, R., Steed, A., Slater, M., Frederick, P., Brooks, J.: Walking > walking-in-place > Flying, in virtual environments. In: Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1999, pp. 359–364. ACM Press/Addison-Wesley Publishing Co, New York (1999)

3. Heintz, M.: Real Walking in Virtual Learning Environments: Beyond the Advantage of Naturalness. In: Cress, U., Dimitrova, V., Specht, M. (eds.) EC-TEL 2009. LNCS, vol. 5794, pp. 584–595. Springer, Heidelberg (2009)
4. Sherman, W.R., Craig, A.B.: Understanding Virtual Reality. Elsevier Science, USA (2003)
5. De Luca, A.: The Motion Control Problem for the CyberCarpet. In: Proceeding of the 2006 IEEE International Conference on Robotics and Automation, pp. 3532–3535. IEEE Press, Florida (2006)
6. CyberWalk Project, <http://www.cyberwalk-project.org>
7. The String Walker Project,
<http://intron.kz.tsukuba.ac.jp/stringwalker/>
8. CirculaFloor,
http://intron.kz.tsukuba.ac.jp/CirculaFloor/CirculaFloor_j.htm
9. Optitrack Motion Capturing System,
<http://www.naturalpoint.com/optitrack>
10. Medved, V.: Measurement of Human Locomotion. CRC Press, New York (2001)
11. Winter, D.A.: Human balance and posture control during standing and walking. *Gait and Posture* 3, 193–214 (1995)
12. Elble, R.J., Moody, C., Leffler, K., Sinha, R.: The Initiation of Normal Walking. *Mov. Disord.* 9(2), 139–146 (1994)
13. Breniere, Y.: When and How Does Steady State Gait Movement Induced from upright posture begin? *Journal of Biomechanics* 19(12), 1035–1040 (1986)
14. Halvorsen, K., Eriksson, M.: Minimal Marker Set for Center of Mass Estimation in Running. *Gait and Posture* 30, 552–555 (2009)
15. Farley, C.T., Ferris, D.P.: Biomechanics of Walking and Running: Center of Mass Movements to Muscle Action. *Exercises and Sport Sciences Reviews* 26, 253–285 (1998)
16. Forsell, C., Halvorsen, K.: A Method for Determining Minimal Sets of Markers for the Estimation of Center of Mass, Linear and Angular Momentum. *Journal of Biomechanics* 42(3), 361–365 (2009)
17. Whittle, M.W.: Gait Analysis: An Introduction. Butterworth-Heinemann, Oxford (2007)

A Survey on Multi-robot Patrolling Algorithms

David Portugal and Rui Rocha

Instituto de Sistemas e Robótica, Department of Electrical and Computer
Engineering, University of Coimbra, 3030-290 Coimbra, Portugal
`{davidbsp, rprocha}@isr.uc.pt`

Abstract. This article presents a survey on cooperative multi-robot patrolling algorithms, which is a recent field of research. Every strategy proposed in the last decade is distinct and is normally based on operational research methods, simple and classic techniques for agent's coordination or alternative, and usually more complex, coordination mechanisms like market-based approaches or reinforcement-learning. The variety of approaches differs in various aspects such as agent type and their decision-making or the coordination and communication mechanisms. Considering the current work concerning the patrolling problem with teams of robots, it is felt that there is still a great potential to take a step forward in the knowledge of this field, approaching well-known limitations in previous works that should be overcome.

Keywords: Multi-Robot Systems; Patrol; Topological Maps and Graph Theory.

1 Introduction

To patrol is “the activity of going around or through an area at regular intervals for security purposes” [1]. In this context, the patrolling task should be performed by multiple mobile robots, which is considered a contemporary area with relevant work presented in the last decade, especially in terms of strategies for coordinating teams of mobile robots. We focus on area patrol, i.e., the activity of going through an area, as opposed to going around an area (perimeter patrol).

Patrolling is a somehow complex multi-robot mission, requiring agents to coordinate their decision-making with the ultimate goal of achieving optimal group performance. It also aims at monitoring and supervising environments, obtaining information, searching for objects and detecting anomalies in order to guard the grounds from intrusion, which involves frequent visits to every point of the infrastructure.

Robotic agents are normally endowed with a metric representation of the environment, which is typically an occupancy grid model, which in turn, is usually abstracted by a simpler, yet precise representation: a topological map (i.e., a graph). Having a graph representation, one can use vertices to represent specific locations and edges to represent the connectivity between those locations. The multi-robot patrolling problem can, thus, be reduced to coordinate robots in order to visit all vertices of the graph ensuring the absence of intruders or other abnormal situations, with respect to a predefined optimization criterion, e.g. the average idleness of the

vertices of the graph. It is consensual that a good strategy should minimize the time lag between visits in strategic places of the environment.

2 Contribution to Sustainability

The major motivation for studying this issue relates to its wide spectrum of applicability and the potential to replace or assist human operators in tedious or dangerous real-life scenarios, such as mine clearing or rescue operations in catastrophic scenarios, easing arduous, tiring and time-consuming tasks and offering the possibility to relieve human beings, enabling them to be occupied in nobler tasks like, for example, monitoring the system from a safe location.

Also, the patrolling problem is very challenging in the context of multi-robot systems, because agents must navigate autonomously, coordinate their actions, be distributed in space, may have communication constraints and must be independent of the number of robots and the environment's dimension. All of these characteristics result in an excellent case study in mobile robotics and conclusions drawn in this field of research may support the development of future approaches not only in the patrolling domain but also in multi-robot systems in general.

3 Pioneer Methods

One of the first works in this field is described in [2] and in more detail in [3], where several architectures of multi-agent patrolling and various evaluation criteria were presented.

Different agent behaviors are employed in the approaches described therein, namely in the agent's perception, which can be reactive (with local information) or cognitive (with access to global information). Also, these architectures differ in the communication mechanism and in the decision of the next vertex to be visited in the topological map.

To analyze the performance of each technique, criteria based on the average and maximum idleness of the vertices were proposed. Random decision algorithms scored the worst results and simple techniques conducted by the vertices' idleness scored close results to the same technique using a centralized coordinator. In general, the best strategy was a local strategy with no communication, based on individual idleness and without a centralized coordinator, called "Conscientious Reactive". Other good results were obtained by "Conscientious Cognitive", which is a similar method; however, agents are no longer reactive, choosing the next vertex to visit on the global graph (instead of their neighborhood).

There are a few weaknesses in this work. Conclusions were drawn based on only two particular environments. Also, unweighted edges were used, meaning that agents travel from one vertex to another in every iteration, independently of the distance between them, which is a rather imprecise simplification. Moreover, the solutions presented are directed to virtual agents in simulation environments, since no real robots were used during experiments.

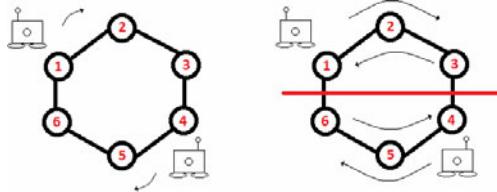


Fig. 1. On the left, an example of a cyclic strategy. On the right, an example of a partition based strategy.

Following the pioneer work of Machado et al., a more complete study was conducted in [4], enriching the representation with weighted graphs, using more and distinct environments, and proposing advanced versions of Machado's architectures to solve the same problem. Essentially, a new path-finding technique was implemented for cognitive agents, based not only in the shortest path between the source and destination vertex but also in the instantaneous idleness of the vertices along the current and goal position. In addition, a decision-making heuristic that considers the idleness of a candidate vertex, as well as the distance to that same vertex, was developed.

The main results show that cognitive architectures have a significant gain when using the new heuristic, if compared to the same technique with the original decision-making process. On the other hand, there was no great benefit for reactive architectures, only in cases when there is high connectivity (several choices of paths between two vertices in the graph) and great variance in the edges' weight. As for pathfinder agents, the performance is always better, especially for graphs with high connectivity. Combining the heuristic decision-making with pathfinder agents and using communication via a central coordinator, one gets the best performing approach in all criteria: Heuristic Pathfinder Cognitive Coordinator.

In [5] the area patrol algorithm developed guarantees that each point in the target area is covered at the same optimal frequency. This is done by computing minimal-costs cyclic patrol paths that visit all points in the target area, i.e. Hamilton cycles. Agents are uniformly distributed along this path and they follow the same patrol route over and over again. One of the key aspects of this strategy is the fact that it is robust, being independent of the number of robots. Uniform frequency of the patrolling task is achieved as long as there is, at least, one robot working properly. A possible disadvantage of this approach is its deterministic nature. An intelligent intruder that apprehends the patrolling scheme may take advantage of the idle time between passages of robots in some points of the area.

Similarly, [6] focuses on two graph-theory centralized planning strategies: cyclic strategies, which are similar to the previously described technique, however it uses a heuristic to compute a TSP¹ cycle; and partitioning strategies, which are approaches that use segmentation of the environment and assign different patrolling regions to each agent. Examples of such strategies are presented in Figure 1.

Both strategies have generally good performance. The first one is better suited for graphs that are highly connected or have large closed paths. The second one is better when graphs have long corridors separating regions. Also, the author explains that very simple strategies, with nearly no communication ability, can achieve impressive results.

¹ TSP stands for the well-known Travelling Salesman Problem (a NP-hard problem).

4 Alternative Methods

In [7], the patrolling task is modeled as a reinforcement learning problem in an attempt to allow automatic adaptation of the agents' strategies to the environment. In summary, agents have a probability of choosing an action from a finite set of actions, having the goal of maximizing a long-term performance criterion, in this case node idleness. Two Reinforcement Learning Techniques, using different communication schemes were implemented and compared to non-adaptive architectures. Although not always scoring the best results, the adaptive solutions are superior to other solutions compared in most of the experiments. The main attractive characteristics in this work is distribution (no centralized communication is assumed) and the adaptive behavior of agents, which can be desirable in this domain.

In [8], a negotiation mechanism is proposed. Agents reveal a scalable and reactive behavior, being able to patrol infrastructures of all sizes and topologies. Also, they need no learning time or path pre-computation.

Each agent acts as a negotiator and receives a set of random graph vertices to patrol. Agents negotiate those vertices using auctions to exchange them with other agents. Aiming to minimize visits to the same node, these agents will try to get a set of nodes in the same region of the graph. Results show that the proposed strategy outperforms most of previously described architectures.

Likewise, [9] also studied cooperative auction systems to solve the patrolling problem. They consider robot's energetic issues and the length of the patrolling routes as performance criteria. However they only experiment in an open space scenario with no obstacles. Nevertheless, despite its weaknesses, the cooperation method among robots has the potential to be used in future works.

A comparative study up to 2004 was presented in [10], which analyzed many different approaches. They observed that the best strategy depends on the topology of the environment and the agents' population size.

Generally, it was concluded that the TSP cyclic approach has the best performance for most cases. However, this architecture will have problems in dynamic environments, large graphs (because of the complexity of a TSP cycle computation in these cases) and graphs containing long edges, due to its predefined nature. Secondly, agents with no communication ability, whose strategies consisted of moving towards the vertex with the highest idleness, performed nearly as well as the most complex algorithm implemented. In general, heuristic-decision agents and reinforcement learning techniques considered have the second best performance, followed by the Negotiation Mechanisms techniques.

The first known patrolling approach, which was focused and implemented on robotic agents, was presented in [11]. Patrolling is seen in a task allocation perspective, where each robot is assigned a different region to visit. Both a reactive and an adaptive approach are described to solve the area patrolling problem, through task data propagation.

Robots send their current state to a centralized system running on a remote computer, through a wireless communication network, to compute the task strength and drive the robot through propagated data. In the experimental setup, robots can estimate their remaining autonomy, thus battery recharges are taken into account and physical interference can occur. The authors claim that efficient patrol is achieved and present interesting properties of adaptability concerning group size and the environment.

In a work with wider scope, [12] presents a motion planning algorithm, which selects effective patrol patterns based on a neural network coverage approach where each robot becomes responsible for a patrol region of the environment. When they operate in patrol mode, robots may update their 3D maps to incorporate possible changes in the environment. If an intruder is detected, some robots will switch their operation mode to threat response scenario and the algorithm is run to successfully respond to the threat, guaranteeing that robots reach the evader in the quickest possible way. The others robots will carry on with the patrol task and replan their trajectories to compensate for those that switched their operation mode.

Recently, swarm intelligence has also been used to tackle the multi-robot patrolling problem as in [13], where a grid-based algorithm is proposed. It relies on the evaporation process of pheromones dropped by agents (an indicator of time passed since the last visit). The agents' behavior is naturally defined by moving towards cells containing less pheromone quantity. Agents only have local perception, and follow paths according to the pheromone quantity in their neighboring cells.

Results show that an approach with global perception is more effective in more complex infrastructures, in terms of idleness. However, it proves to be twice as costly, in terms of computational complexity. Due to the marking of the environment, the system self-organizes and an effective patrolling behavior emerges.

5 Recent Studies

In a very thorough study, [14] explores the concept of unpredictability in the multi-agent area patrol task. In this context, intruders will not have access to patrolling trajectory information. Metrics are presented in terms of probability of catching intruders with different intelligence. The authors evaluate six different algorithms. Two are purely random (one locally and one globally), one is a deterministic TSP-based solution and three are based on the TSP solution together with local random nuances to create unpredictability in the trajectories.

Intensive simulation results, using a large set of graphs, made evident some distinct facts. Partitioning schemes are more effective against random attackers; however, non-partitioning schemes perform better when the attacker has some level of intelligence. With random attackers or attackers with limited information, the deterministic TSP algorithm was the best solution found, because it covers all the nodes with the minimal time needed. The algorithm that performed better against statistical intruders was an unpredictable variant of the TSP cycle, thus confirming the importance of unpredictability in the patrolling task. Random-based strategies although being very unpredictable, have very high worst idleness values, which makes them generally useless for the patrolling problem.

Patrolling related issues were also studied in [15] like map representation, graph extraction, surveillance, pursuit-evasion, coverage, navigation and exploration strategies. By analyzing approaches, not limited exclusively to patrolling works, an original, scalable, centralized and efficient algorithm was presented, called Multilevel Subgraph Patrolling (MSP) Algorithm, which is also described in [16].

The MSP algorithm is a multilevel partitioning algorithm that assigns different regions (subgraphs) to each mobile agent. The algorithm deals with effectively

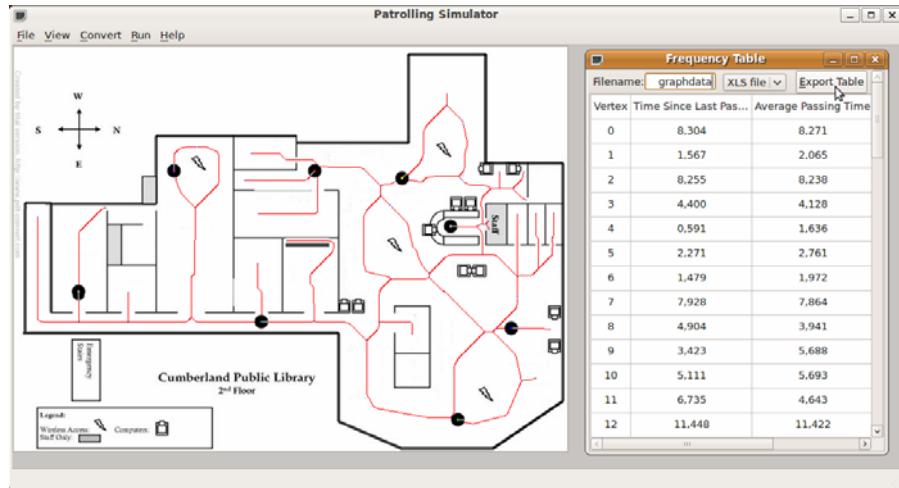


Fig. 2. MSP running on the patrolling simulator window [15]

computing paths for every robot using a classical algorithm for Euler cycles and various heuristics for Hamiltonian cycles, non-Hamiltonian Cycles and Longest paths. The algorithm was compared to a cyclic algorithm, like the one presented in [5]. The MSP Algorithm scored slightly better results in half of the cases and obtained slightly worse results in the other half. Given that cyclic algorithms are well-known for their performance in terms of visiting frequency, these results are very optimistic and confirm the flexible, scalable and high performance nature of the approach, which also benefits from being non-redundant and does not need inter-agent communication. However, like the cyclic algorithm, MSP is deterministic. Nevertheless, it is much more difficult for an evader to attack in this case, because it would need to keep track of every single robot local patrolling path. Following this work, four different genetic algorithms for approximating the longest path in a graph were also presented in [17].

In the next page, Table 1 was built to summarize the main characteristics of the architectures described in the previous sections.

6 Existing Limitations and Future Work

The analysis made allows us to conclude that several drawbacks, common to most strategies, remain. Among others, future work will focus on overcoming weaknesses such as: the absence of studies on scalability, flexibility, resource utilization, interference, communication load or workload among robots when performing the patrolling task; the lack of experimental work using teams of robots in real scenarios; simplifications and unfeasibility of simulation approaches towards real life experiments; lack of comparisons of the actual time spent between different strategies in their patrolling cycles; lack of diversity and classification of the environments tested and the deterministic nature of many centralized approaches.

Table 1. Summary of the main architectures analyzed

Proposed Strategy	Type/Perception	Communication	Coordination	Decision-Making
Conscientious Reactive [3]	Reactive / Local	None	Emergent	Local Idleness-based
Conscientious Cognitive [3]	Cognitive / Global	None	Emergent	Global Idleness-based
Idleness Coordinator Monitored [3]	Cognitive / Global	Coordinator Messages	Centralized	Idleness-based with Monitoring
Heuristic Pathfinder Cognitive Coordinated [4]	Cognitive / Global	Coordinator Messages	Centralized	Heuristic (Idleness + Distance) with Path-finding technique
Untitled #1 [5]	Cognitive / Global	Coordinator Trajectory Cycle	Centralized	Hamilton Cycle Computation
Cyclic Approach [6]	Cognitive / Global	Coordinator Trajectory Cycle	Centralized	TSP Heuristic Calculation
Partitioning Approach [6]	Cognitive / Global	Coordinator Trajectory Cycle	Centralized	TSP Heuristic inside each region
Gray-Box Learner Agent [7]	Reactive / Local	Flags	Emergent + Adaptive	Idleness-based Reinforcement Learning with Monitoring
Bidder Agent [8]	Cognitive / Global	Bidding Messages	Auctions	Self-Interested Idleness-based
Sequential Single-Item Auctions [9]	Cognitive / Global	Bidding Messages	Auctions	Minimize the maximum patrol path
Heuristic Pathfinder Two-Shot Bidder [10]	Cognitive / Global	Bidding Messages	Two-Shot Auctions	Heuristic (Idleness + Distance) with Path-finding technique
Untitled #2 [11]	Reactive / Local	Task Propagation Messages	Centralized + Adaptive	Task strength Idleness-based measure
Untitled #3 [12]	Cognitive / Global	Coordinator Messages	Centralized	Neural network coverage approach inside each region
EVAP [13]	Reactive / Local	Flags	Emergent	Local Idleness-based
CLInG [13]	Reactive / Local	Flags	Emergent	Idleness-based with diffusion of information
TSP rank of solutions [14]	Cognitive / Global	Coordinator Messages	Centralized	Queue of TSP sub-optimal solutions (Unpredictable)
MSP [16]	Cognitive / Global	Coordinator Trajectory Cycle	Centralized	Operation research algorithms inside each region

To materialize this, we intend to create a formal mathematical model for the multi-robot patrolling problem by analyzing and comparing the performance of different approaches, through realistic simulations, focusing especially on how these teams scale and extracting the important variables for this problem, so as to create an automatic estimation tool to dimension the team for a patrolling task. In addition, we intended to develop a new distributed, non-deterministic and cooperative strategy, which will be tested firstly by simulations in a wide variety of environments and secondly using a team of mobile robots in real-world scenarios.

Acknowledgments

This work was financially supported by a PhD grant (SFRH/BD/64426/2009) from the Portuguese Foundation for Science and Technology (FCT) and the Institute of Systems and Robotics (ISR-Coimbra) also under regular funding by FCT.

References

1. Webster's Online Dictionary (November 2010), <http://www.websters-online-dictionary.org>
2. Machado, A., Ramalho, G., Zucker, J., Drogoul, A.: Multi-Agent Patrolling: an Empirical Analysis of Alternative Architectures. In: Sichman, J.S., Bousquet, F., Davidsson, P. (eds.) MABS 2002. LNCS (LNAI), vol. 2581, pp. 155–170. Springer, Heidelberg (2003)
3. Machado, A.: Patrulha Multiagente: Uma Análise Empírica e Sistemática. M.Sc. Thesis, Centro de Informática, Univ. Federal de Pernambuco, Brasil (2002) (in Portuguese)
4. Almeida, A.: Patrulhamento Multiagente em Grafos com Pesos. M.Sc. Thesis, Centro de Informática, Univ. Federal de Pernambuco, Recife, Brasil (2003) (in Portuguese)
5. Elmaliah, Y., Agmon, N., Kaminka, G.: Multi-Robot Area Patrol under Frequency Constraints. In: Int. Conf. on Robotics and Automation, Rome, Italy, pp. 385–390 (2007)
6. Chevaleyre, Y.: Theoretical Analysis of the Multi-agent Patrolling Problem. In: Proc. of the Int. Conf. On Intelligent Agent Technology, Beijing, China, pp. 302–308 (2004)
7. Santana, H., Ramalho, G., Corruble, V., Ratitch, B.: Multi-Agent Patrolling with Reinforcement Learning. In: Proc. of the Third Int. Joint Conf. on Autonomous Agents and Multiagent Systems, New York, vol. 3, pp. 1122–1129 (2004)
8. Menezes, T., Tedesco, P., Ramalho, G.: Negotiator Agents for the Patrolling Task. In: Sichman, J.S., Coelho, H., Rezende, S.O. (eds.) IBERAMIA 2006 and SBIA 2006. LNCS (LNAI), vol. 4140, pp. 48–57. Springer, Heidelberg (2006)
9. Hwang, K., Lin, J., Huang, H.: Cooperative Patrol Planning of Multi-Robot Systems by a Competitive Auction System. In: Int. Joint Conf., Fukuoka, Japan, August 18-21 (2009)
10. Almeida, A., Ramalho, G., Sanana, H., Tedesco, P., Menezes, T., Corruble, V., Chevaleyre, Y.: Recent Advances on Multi-Agent Patrolling. In: Bazzan, A.L.C., Labidi, S. (eds.) SBIA 2004. LNCS (LNAI), vol. 3171, pp. 474–483. Springer, Heidelberg (2004)
11. Sempé, F., Drogoul, A.: Adaptive Patrol for a Group of Robots. In: Proc. of the Int. Conf. on Intelligent Robots and Systems, Las Vegas, Nevada (October 2003)
12. Guo, Y., Parker, L., Madhavan, R.: 9 Collaborative Robots for Infrastructure Security Applications. In: Studies in Computational Intelligence (SCI), April 22, v 2007, vol. 50, pp. 185–200. Springer, Heidelberg (2007)
13. Chu, H., Glad, A., Simonin, O., Sempé, F., Drogoul, A., Charpillet, F.: Swarm Approaches for the Patrolling Problem, Information Propagation vs. Pheromone Evaporation. In: Int. Conf. on Tools with Art. Intelligence, France, vol. 1, pp. 442–449 (2007)
14. Sak, T., Wainer, J., Goldenstein, S.: Probabilistic Multiagent Patrolling. In: Proc. of the Brazilian Symposium on Artificial Intelligence, Salvador, Bahia, Brazil (2008)
15. Portugal, D.: RoboCops: A Study of Coordination Algorithms for Autonomous Mobile Robots in Patrolling Missions. Msc. Dissertation, Faculty of Science and Technology, University of Coimbra, Portugal (September 2009)
16. Portugal, D., Rocha, R.: MSP Algorithm: Multi-Robot Patrolling based on Territory Allocation using Balanced Graph Partitioning. In: Proc. of Symposium on Applied Computing (SAC 2010), Sierre, Switzerland, March 22-26, 2010, pp. 1271–1276 (2010)
17. Portugal, D., Henggeler Antunes, C., Rocha, R.: A Study of Genetic Algorithms for Approximating the Longest Path in Generic Graphs. In: Proc. 2010 IEEE Int. Conf. on Systems, Istanbul, Turkey, October 10-13 (2010)

Autonomous Planning Framework for Distributed Multiagent Robotic Systems

Marko Švaco, Bojan Šekoranja, and Bojan Jerbić

University of Zagreb, Faculty of Mechanical Engineering and Naval Architecture,
Department of Robotics and Production System Automation

Ivana Lučića 5, 10000 Zagreb, Croatia

{marko.svaco, bojan.sekoranja, bojan.jerbic}@fsb.hr

Abstract. In this paper a creative action planning algorithm (CAPA) is presented for solving multiagent planning problems and task allocation. The distributed multiagent system taken in consideration is a system of m autonomous agents. Agents workspace contains simplified blocks which form different space structures. By employing the planning algorithm and through interaction agents allocate tasks which they execute in order to assemble the required space structure. The planning algorithm is based on an inductive engine. From a given set of objects which can differ from the initial set agents need to reach a solution in the anticipated search space. A multiagent framework for autonomous planning is developed and implemented on an actual robotic system consisting of three 6 DOF industrial robots.

Keywords: Distributed robotic system, autonomous planning, multiagent system, assembly, industrial robotics.

1 Introduction

Substantial research and development is conducted to multiagent robotics; particularly in the fields such as service, humanoid or mobile robotics, but industrial robotics is still based on traditional postulates. Real flexibility and adaptivity to changes are shortcomings in today's industrial assembly and handling robotic applications and are issues that need to be addressed. Distributed multiagent robotics is a system based on human behavior patterns. When complex tasks arise humans are much more efficient when working in groups: they exhibit more axis of freedom, more data can be handled and they delegate particular tasks to individual agents.

Research concentrated around humanoid robotics ([1]-[2]) is developing rapidly. Dual arm configuration highly sophisticated perceptive mechanisms, human like motions enable robots to recreate human motion and work patterns. Major drawback of those kinematical structures is very low repeatability and precision primarily needed in industrial systems. For assembly and handling tasks which usually have high precision and repeatability demands industrial robots are necessary. Nowadays the most flexible industrial robots have 6 or 7 [3] degrees of freedom (DOF) without the end effector (gripper) which usually has 1 additional DOF. One human arm (with the hand) has 27 DOF [4]. The flexibility of a robotic arm is quite limited in

comparison to a human operator. Implementing two or more robots with own controllers that communicate each with other, a certain multiagent concept can be achieved. The whole system will be orchestrated and will be able to perform more demanding operations. Each controller running its own actuator unit should be an agent with defined level of autonomy. In such systems the multiagent control appears as the main issue.

In this paper a creative action planning algorithm (CAPA) for application in multi agent robotic systems is presented. One of the main goals is constructing a universal planning framework which can be implemented on various types of industrial robots and tasks.

Related works [5], [6] incorporating multiagent planning on similar tasks are virtual applications and cannot be easily implemented on real industrial systems. The approaches are primarily intended for autonomous planning done by multiple agents who cannot collide, are of infinite small dimension and share the same computational time domain. The developed CAPA and the distributed multiagent system (MAS) operate in a real world environment bounded by rules and limitations. The approach discussed in this paper is intended to show that some assembly and handling tasks can be done in close collaboration among agents to gain flexibility and increase overall system productivity.

2 Contribution to Sustainability

Robotics and in particular industrial robotics have always been a part of a central planning system. Agents (robots, machines) controlled by own computers are somehow subordinated to a central system controller [7]. Therefore they exhibit very low level of autonomy and in most cases do pre-programmed actions not being able to cope with uncertainties in the system and the environment. Uncertainties may vary from production quantities to failures of equipment or other agents, etc.

It is suggested that some handling and assembly industry tasks can be accomplished by interaction between agents (primarily industrial robots) in the system. Accordingly some level of autonomy must be introduced.

Production in recent years has switched from high quantity standardized products to lower quantities of customized products so demands from assembly systems have grown. Traditional approach with a centralized architecture and strict delegation of tasks needs to be replaced. Introducing a multiagent configuration and autonomous planning approach could be proven as a valuable addition. For an assembly system it implies that agents (robots and machines) before assembling need to generate a plan that best suits the current state and requirements of the system. After deriving consensus agents begin assembling the structure (product) constantly communicating and exchanging relevant information and data. In industrial assembly systems this is a novice approach and it has numerous benefits when implemented: it leads to increased flexibility and adaptivity to unexpected changes and uncertainties in the system, i.e. responsiveness [8]. The system becomes insensitive to number of agents (robots) and new assembly tasks can be resolved with less effort. Clearly, this approach is not suited for all products but it can be implemented on a variety of industrial examples. Development of such an industrial system scheme is beyond the scope of this paper and will be considered for further research. In this work an initial version

of the planning framework is presented. The framework is implemented on an actual system consisting of three 6 DOF robots.

3 The Multiagent System

3.1 System Formulation

The multiagent system consists of m autonomous agents a_l ($l = 1 \dots m$). Their workspace contains simplified blocks with respect to a global Cartesian coordinate system K . Agents workspace $W(a_i, b_{j,k})$ contains blocks which form different space structures, where $b_{j,k}$ represents j^{th} block of k^{th} type. Each block has certain properties which agents perceive from their workspace: *size (type)* of a building block $T(b_{j,k}) = \{1, 2, 3 \dots\}$ and *Cartesian position and orientation* in workspace: $P(b_{j,k}) = \{x, y, r\}$. All blocks have the same width and height (single unit) but their length can vary and can be one, two, three, etc. unit lengths. That results with flexibility so building blocks can be supplemented with each other i.e. block with two unit lengths can be replaced with two blocks of single unit length and vice versa. Each agent is defined as an autonomous, self-aware entity with limited knowledge of the global workspace [9] and with some cognition of other agents. It has a separate processing unit, actuators, vision system for acquiring information from its environment, force and torque sensors for haptic feedback and other interfaces. A space function $F(a_l)$ is defined to determine the consumed space by an agent a_l , $F=(x_1, y_1, x_2, y_2, r, t)$ in time t . The first pair of Cartesian coordinates depicts the first vertex of a rectangle which bounds the agent and the second pair depicts the second vertex respectively. Rotation angle r is defined with respect to the origin point of the coordinate system K .

The MAS is insensitive to dynamic changes in number of agents. Impact is lower system flexibility and longer times for achieving final goals when agents are excluded from the system.

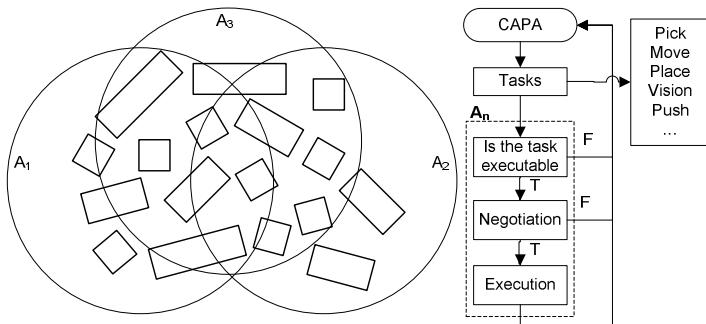


Fig. 1. Agent workspace and task allocation scheme

3.2 Structures and Decision Making

Agents' tasks are recreating structures which are defined as a final form put together from various objects with defined relationships. A structure is determined by

interrelations and arrangement of objects $b_{j,k}$ into a complex entity. Structure $S = \{R_i b_{j,k}\}$ is a set of relations R_i ($i = 1 \dots m-1$) between objects $(b_{j,k}, j = 1 \dots n, k = 1 \dots u)$.

The MAS has properties of a market organization type [10], [11] where agents bid [12] for given resources (blocks) in their workspace (Fig. 1). Time schedules need to be negotiated when areas of interest in the global workspace are not occupied.

Global goal G is the required structure that must be assembled from available elements following the given set S . An example of a structure is illustrated in Fig. 2 a). After observing a structure and finding relations agents are given an arbitrary set of work pieces (blocks) as depicted in Fig. 2. (b). Using those elements a plan of actions is generated for assembling the initial structure. Possible solutions are presented in Fig. 2. (c₁) - (c₃). A set of rules and propositions for agent behavior is given in a cognition base (CB):

- Mathematical rules for structure sets
- Agents capabilities
- Grasping rules and limitations
- Object properties
- Agent workspace
- Vision system patterns database
- Force and torque sensor threshold values

If a simple structure with limited number of building blocks is presented to the agents (Fig. 2 a) there might be only one or few feasible solutions (sequence of steps). If more complex structures are presented (as shown in Fig. 4. a) a variety of feasible solutions might be possible.

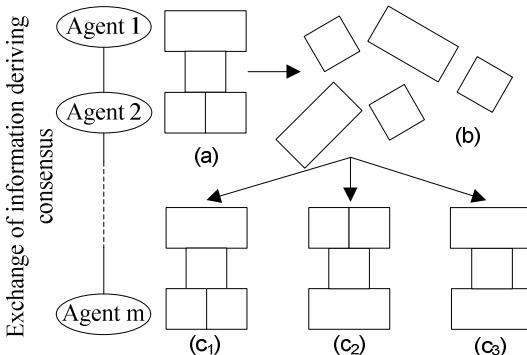


Fig. 2. Simple task for the multiagent system

Top down disassembling or bottom up assembling the structure can define a sequence of steps for the MAS. The CAPA utilizes a bottom up principle where from a provided set of objects $\{b_1 \dots b_{p,r}\}$, which can differ from the initial set $\{b_1 \dots b_{n,q}\}$ agents need to reach a solution in the given search space. Depending on the CB information agents can make decisions whether the desirable objectives can be performed in accordance to proposed restrictions and limitations. Implementing an

iterative algorithm a solution can be found as shown in Fig. 3. Branches represent solution sets and each branch leads to one solution. If finding a solution in one solution set isn't possible, the system takes one step back and explores other options until it finds a valid one.

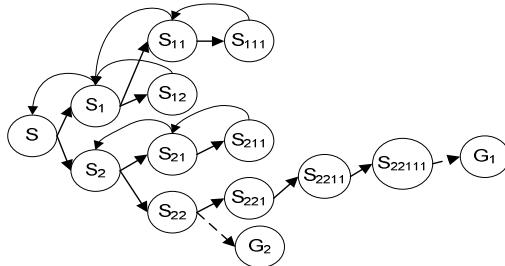


Fig. 3. Possible solution sets (sequence of steps) for a represented structure

Each agent attains a unified set of sub goals g_i , which fulfill the global goal G . Execution of sub goals (tasks) can be done synchronous or asynchronous giving the space functions F of the agents. A resource function C is defined as a measure of resource and time consumption. $C(a_i, b_{j,k}, e)$ is a function of agents' a_i position, specifications of a building block $b_{j,k}$ (size and position in global workspace) and the position e where that block is planned to be moved.

3.3 Operators

By utilizing operators agents construct a sequence of actions for accomplishing each sub goal. By consecutively achieving all sub goals the global goal G is fulfilled and the agents await further tasks. The basic operators are:

- *Pick* (b_i, gr_k) – agent picks up a block b_i with a grasping method gr_k
- *Move* ($p_1, \dots, p_r, t_1, \dots, t_r$) – agent moves in the global workspace from point p_1 to point p_r through $r-2$ interpolation points with motion specification t_r defined for each point.
- *Place* ($b_{j,k}$) - agent places a block $b_{j,k}$
- *Vision* – vision operator is used for identifying objects and their coordinates in c
- *Push* (f, d, s) – agent uses force/torque sensor for auxiliary action of pushing an object with force/torque threshold t in vector direction d for s units
- *Force* – used for positioning correction

The vision operator utilizes the cognition base and solves problems of identifying work objects and associated data. Therefore vision processes have to be very stable and work in constantly changing light and scenery conditions [13]. A fix to this problem is to utilize algorithms that can change the exposition of the image acquisition process. This can be done through a way of search patterns. Few images are taken at different camera settings and the one where familiar objects are recognized is used as

reference. The downside is increase of image acquisition and processing time. If light and scenery conditions can vary this is a necessity due to the high level sensitivity of precision vision applications.

Furthermore if there is a need for even higher precision beyond capabilities of vision systems agents can use very sensitive force sensors. This method improves accuracy and corrects the pick&place positions. To determine which method to apply agents rely on vision identification of objects. Once the object is identified agents can decide which method of force correction to apply. In example they can successfully insert a shaft or a square object into adequate holes if their original position was slightly off the required one. Furthermore the force sensor allows an agent to correct larger errors. A controlled search pattern is used with a very low force not to damage the objects. Finding the adequate insertion position completes the process.

3.4 Global System Approach

The starting point of every assembling process is perception of the agent's environment. Each agent uses vision systems to acquire information from a portion of the global workspace and forwards it to the planning agent. From the global information set the planning agent extracts relations between objects forming the initial structure. The same principle is applied for the random set of work objects. Regarding the initial structure and available elements the planning agent decomposes the global goal into tasks which can be performed by an individual agent. Task priorities are also taken into consideration where some tasks are conditioned to be executed before others. After this initial process each agent bids for a task. Through comparing resource functions agents submit the task to the optimal candidate. Idle agents repeat this process and acquire free tasks. When processing a task an agent sends data regarding the consumed space through the F function for collision avoidance. After all task are allocated and executed agents inspect the reassembled structure.

4 Implementation

The CAPA has been tested to provide solutions for a structure such as the one shown in Fig. 4. When multiple solutions are possible the MAS executes the one where $\sum C$ in the entire solution set is minimal. Currently only two dimensional structures (R^3) are being solved but their solutions due to use of real world objects has to be three dimensional (R^4). The planning algorithm was tested on a virtual model of the multi-agent robotic system (Fig. 5. a). This was done for safety reasons (primarily collision) and the ability to test and debug the algorithm in parallel on multiple computers. After satisfactory computational results the algorithm and the entire framework have been implemented on an actual robotic system – Fig. 5. (b).

The first problem which emerged was sharing of agent workspace. In order to work on the same task, assembling the same structure, spatial relations need to be taken into consideration. Agents were calibrated using calibration tools and visual applications.

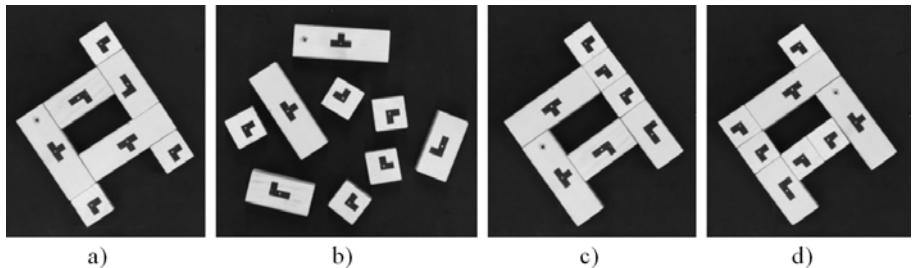


Fig. 4. a) Initial space structure b) Randomly scattered building blocks c), d) Space structures assembled by the agents

This creates relations with respect to agent positions (three translations) and rotations (three angular displacements); introducing a common global workspace (K). A problem that resulted from the decentralized multiagent architecture was sharing and synchronizing agent time domains. This didn't introduce an issue while tests were conducted on a computer where all agents used the same *CPU clock*. Adjustments have been done using handshaking with digital signals and through TCP/IP communication which allowed coordinated task execution.

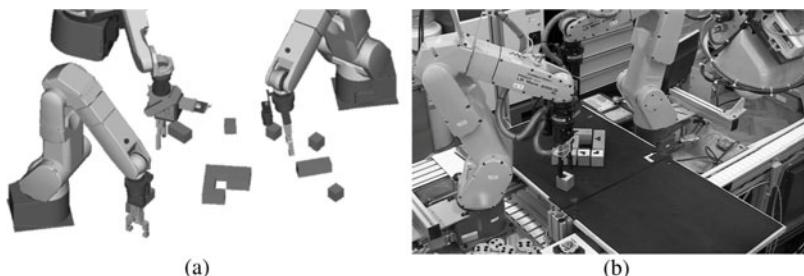


Fig. 5. (a) virtual representation of the multiagent robotics system (b) real agents

Collision detection was an issue that needed to be addressed. Currently there are no algorithms to solve real time agent collision or they exist but with limitations. Collision between two agents with kinematic chains of 3 DOF can be solved in a definite period of time [14]. For the reason of limited computational power and the collision detection not being the centre of this research the function (F) described in chapter 3 was used.

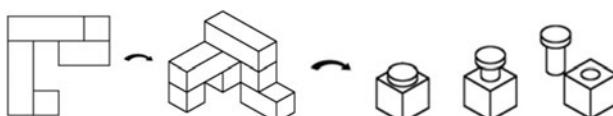


Fig. 6. Planning of 3D structures, following research direction

5 Conclusion and Further Research

The approach presented in this paper ensures higher system robustness regarding decentralized task execution. In future applications agents would decide how to best solve a new problem – which agent has the most adequate tools and how can the rest of them assist etc. Making agents mobile and giving them the ability to exchange tool heads introduces even a greater level of flexibility to the system. This results in more cost optimized solutions. Further generalization will be introduced where agents will be able to autonomously distinguish and solve entirely new problems as illustrated in Fig. 6. First step is implementation of reassembling 3D structures. For that purpose the CB will need to comprise rules regarding “laws of gravity” and etc. Further research will be concentrated on introduction of new objects to the MAS. Taking into consideration the CB (known similar objects) agents will be able to find or construct grasping methods whether they can do it individually or assisted by other agents.

References

1. Akachi, K., Kaneko, K., Kanehira, N., Ota, S., Miyamori, G., Hirata, M., Kajita, S., Kanehiro, F.: Development of humanoid robot hrp-3. In: 5th IEEE/RAS International Conference on Humanoid Robots, pp. 50–55 (2005)
2. Park, I.W., Kim, J.Y., Lee, J., Oh, J.H.: Mechanical design of humanoid robot platform khr-3 (kaist humanoid robot-3: Hubo). In: IEEE/RAS International Conference on Humanoid Robots, pp. 321–326 (2005)
3. Bischoff, R.: From research to products: The development of the KUKA Light-Weight Robot. In: 40th International Symposium on Robotics, Barcelona, Spain (2009)
4. Agur, A.M.R., Lee, M.J.: Grant's Atlas of Anatomy. Lippincott Williams and Wilkins, Baltimore (1999)
5. Sycara, K., Roth, S., Sadeh, N., Fox, M.: Distributed Constrained Heuristic Search. *IEEE Trans. on System, Man and Cybernetics*, 1446–1461 (1991)
6. Ephrati, E., Rosenschein, J.: Divide and conquer in multiagent planning. In: Proc. of the 12th National Conference on AI, pp. 375–380. AAAI, Seattle (1994)
7. Tang, H.P., Wong, T.N.: Reactive multi-agent system for assembly cell control. *Robotics and Computer-Integrated Manufacturing* 21, 87–98 (2005)
8. Seilonen, I., Pirttioja, T., Koskinen, K.: Extending process automation systems with multi-agent techniques. *Engineering Applications of Artificial Intelligence* 22 (2009)
9. Schumacher, M.: Objective Coordination in Multi-Agent System Engineering. Springer, New York (2001)
10. Sandholm, T.: An Implementation of the Contract Net Protocol Based on Marginal Cost Calculations. In: Proc. of the 11th Conference on AI, pp. 256–262 (1993)
11. Shoham, Y., Leyton-Brown, K.: Multiagent Systems: algorithmic, game-theoretic and logical foundations. Cambridge Uni. Press, New York (2009)
12. Hsieh, F.-S.: Analysis of contract net in multi-agent sys. *Automatica* 42, 733–740 (2006)
13. Stipancic, T., Jerbic, B.: Self-adaptive Vision System. In: Camarinha-Matos, L.M., Pereira, P., Ribeiro, L. (eds.) DoCEIS 2010. IFIP Advances in Information and Communication Technology, vol. 314, pp. 195–202. Springer, Heidelberg (2010)
14. Curkovic, P., Jerbic, B.: Dual-Arm Robot Motion Planning Based on Cooperative Coevolution. In: Camarinha-Matos, L.M., Pereira, P., Ribeiro, L. (eds.) DoCEIS 2010. IFIP AICT, vol. 314, pp. 169–178. Springer, Heidelberg (2010)

Controlling a Robotic Arm by Brainwaves and Eye Movement

Cristian-Cezar Postelnicu¹, Doru Talaba², and Madalina-Ioana Toma¹

^{1,2} Transilvania University of Brasov, Romania, Faculty of Mechanical Engineering,

Department of Product Design and Robotics

{cristian-cezar.postelnicu,talaba,madalina-ioana.toma}@unitbv.ro

Abstract. This paper proposes two paradigms for controlling a robotic arm by integrating Electrooculography (EOG) and Electroencephalography (EEG) recording techniques. The purpose of our study is to develop a feasible paradigm for helping disabled persons with their every-day needs. Using EOG, the robotic arm is placed at a desired location and, by EEG, the end-effector is controlled for grasping the object from the selected location. Simple algorithms were implemented for detecting electrophysiological signals like eye saccades, blinking and eye closure events. Preliminary results of this study are presented and compared.

Keywords: electrooculography, electroencephalography, eye movement, robotic arm, brain computer interface.

1 Introduction

In the European Union, there are about 37 million disabled people [1]. For disabled people with severe neuromuscular disorders such as brainstem stroke, brain or spinal cord injury, cerebral palsy, multiple sclerosis or amyotrophic lateral sclerosis (ALS), we must provide basic communication capabilities in order to give them the possibility to express themselves [2]. Two main solutions have been developed over time: Brain Computer Interface (BCI) systems and EOG based systems.

A BCI is a non-muscular communication channel that enables a person to send commands and messages to an automated system such as a robot or prosthesis, by means of his brain activity [2] and [3]. BCI systems are used in numerous applications such as: speller applications [4], [5] and [8], controlling a wheelchair [6], prosthesis [9] or a cursor on a screen [10], and also for multimedia [11] and virtual reality [12]. One of the most important features in a BCI system is represented by acquisition. The most spread acquisition technique is EEG, and it represents a cheap and portable solution for acquisition. The EEG technique assumes brainwaves recording by electrodes attached to the subject's scalp. EEG signals present low level amplitudes in the order of microvolts and frequency range from 1 Hz up to 100 Hz. Specific features are extracted and associated with different states of patient brain activity, and further with commands for developed applications.

EOG is a technique for measuring the resting potential of the retina by analyzing the surrounding muscles. The eye can be represented as a dipole. The potential in the

eye can be determined by measuring the voltages from the electrodes placed around the eye, eyegaze changes or eye blinks causing potential variations (electrodes placement presented in Fig. 1). In the field of rehabilitation, this technique was used for applications such as virtual keyboard [7], control of a wheelchair [13] or for commanding the hand grippers of a robot[14].

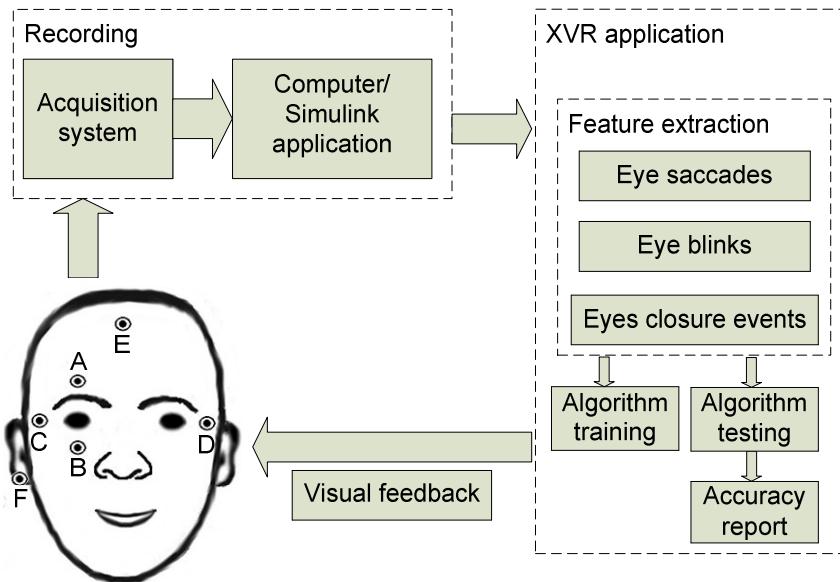


Fig. 1. System architecture and EOG electrodes placement

The EOG signals amplitude values vary from 50 to $3500\mu\text{V}$ [13]. Gaze angles vary linear for $\pm 30^\circ$. Amplitude of biopotentials recorded from patients varies widely even when similar recording conditions are achieved; thus we need a special recognition algorithm for identifying relevant parameters from recorded signals. Different types of signals have been identified and used in applications: blinks and saccadic movements.

In this paper, we focus on presenting and analyzing the results of two developed paradigms for controlling a robotic arm by integrating EOG and EEG signals. Experimental paradigms are presented in the “EOG-based Control Paradigm” and “EOG-EEG-based Control Paradigm” sections. Also, a real-time pattern recognition algorithm for eye saccades and blinking is presented. We conclude by presenting further improvements of current developed paradigms and further phases of our project.

2 Contribution to Sustainability

Our project’s aim is to develop a feasible solution for helping disabled people. In this paper, we focused on a comparison between two possible interaction paradigms. This

paper presents the results of preliminary tests, which are promising for the further development of our project.

A fully functional system will be implemented, thus the user (a disabled person) will have the opportunity to control a robotic system.. A robotic arm will be attached to a wheelchair that is in fact a mobile robot; thus the user will control the robotic arm in order to get help, and also will control the wheelchair navigation. This is going to be tested with the car simulator developed within the project IREAL that has the capability to fully simulate the navigation for training purposes (see Fig. 1). By means of biopotentials, the user will control his wheelchair and will choose a desired location (an object that is found at that location) for positioning the robotic arm's gripper. After selecting a desired object, the user can control the robotic arm in order to manipulate the object, e.g. to feed himself.

Based on results presented in this paper we will decide on a paradigm for our future implemented system. Fatigue is an important factor in developing a feasible system, and one of the two paradigms proposed seems to reduce it when the system is used.

3 Developed System

Helping disabled people is one of the most important research fields nowadays. Two paradigms for controlling a robotic arm are presented in this paper. Although disabled people cannot use their normal output pathways (speaking, moving their limbs) to interact with other people, they can still send us messages by using special pathways. The former paradigm is based exclusively on EOG signals, whereas the latter combines EOG and EEG signals.

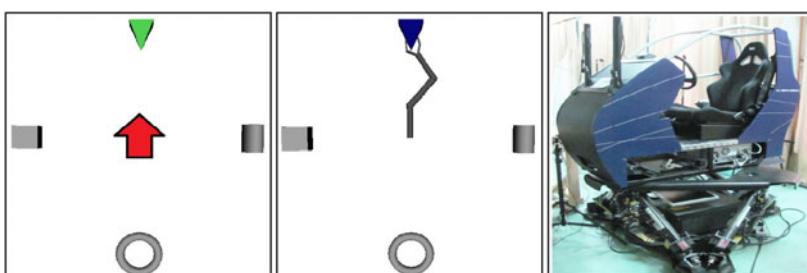


Fig. 2. Visual feedback (four objects are drawn). Left – a successful trial of selecting top object; middle – a successful trial of robotic arm's gripper closure; right – car simulator.

3.1 System Architecture

Our implemented solution consist in recording EEG and EOG patient's biopotentials, and using them for selecting in a synchronous pseudo-randomly manner between four objects placed at fixed locations (see Fig. 2). This preliminary phase of our main project is based on a virtual reality application developed using XVR software [16].

The recording block represents an interface between physical device, used for signal acquisition, and computer, used for signal processing (see Fig. 1). A Simulink application was implemented for acquiring signals from device and forwarding them to the XVR application. Recorded signals were also stored in files for further analysis.

The XVR block represents a developed application for processing acquired signals, identifying relevant parameters, automatically training developed algorithms and testing the proposed control paradigms (see Fig. 1). Also, this application sends a visual feedback for the user. Many previous studies proved that visual feedback enhances the user's attention for the tested application.

3.2 Experimental Setup

For this study, it was used a g.USBamp system from Guger Technologies [17]. This system is a multimodal amplifier for electrophysiological signals like EEG, EOG, ECG (Electrocardiography – recording of the electrical activity of the heart) and EMG (Electromyography – recording of the electrical activity produced by skeletal muscles). Electrodes are placed in pairs (a bipolar recording for EOG signals, see Fig. 1): horizontal movements are detected by C-D pair and vertical movements by A-B pair. Electrode E is placed on the subject's forehead, it represents system ground, and electrode F is placed on the right earlobe and represents the system reference. All electrodes were placed on the subject's skin using a conductive gel. Signals were sampled at 256 Hz and were bandpass filtered between 0.5 Hz and 30 Hz, also an additional 50 Hz notch filter was enabled to suppress line noise.

The user was sitting in front of a computer monitor at around 50 cm. Visual stimuli (objects) were drawn on the monitor at around ± 20 degrees on horizontal and ± 15 degrees on vertical with respect to the centre of the monitor (see Fig. 2). Thus, the drawn objects were selected by eye saccades. When the indicated object was selected, the gripper attached to robotic arm was placed at that location and the user was instructed to close the gripper. In the former solution, the user was able to close the gripper using a double blink action, whereas in the latter solution he could close the gripper by closing his eyes for a short period of time.

The implemented solution consists of two different steps. The former step is represented by training of our pattern recognition algorithm. First, the user is instructed to execute the instructions that appear on the monitor: "Left", "Right", "Top", "Bottom" (these commands were drawn as arrows and were executed by eye saccades, see Fig. 2), "Double blink" and "Closed eyes". This step was also useful for training the subject in using the application. Pattern recognition algorithms for eye saccades movement identification, double blinking and eyes closing events identification were implemented. The latter step of our application consists in testing the implemented algorithms, in terms of accuracy, by presenting at fixed time steps visual instructions for subjects.

3.3 Pattern Recognition Algorithm

For eye movement events, a real-time pattern recognition algorithm was implemented. The algorithm recognizes values for amplitude and width during a requested action. During the training phase, for eye saccades the algorithm identifies

the average of the signals and then compares it with each sample in order to determine the signal's maximum amplitude and width during each event separately. The lowest and highest absolute values were rejected from recorded parameters for each event separately; other values were averaged. The width was determined by selecting continuous values exceeding with at least 25% the difference between current maximum amplitude and signal's average. For double blink events were identified two sets of values, one set for each blink.

A successful trial is considered when values for current requested action (left, right, up, down or double blink), in terms of width and amplitude, have a maximum range of $\pm 20\%$ considering difference between identified values during training phase and current identified values. This percentage was considered after evaluating a set of preliminary recordings. The algorithm works in a real-time manner; it checks every sample and compares current values with previously identified values. A valid trial is considered if the action is executed in a maximum time interval of 2 seconds.

For EEG paradigm the algorithm extracted a maximum average of recorded signals during eye closure events; this value was calculated during training phase. During the testing phase the algorithm was searching for a variation of at least 80% from difference between identified average and average during the resting phase, and also for a minimum duration time of 600ms.

3.4 EOG-Based Control Paradigm

This paradigm assumes eye saccades for object selection and double eye blinking events for gripper closure. An eye saccade represents a potential variation of around $20\mu\text{V}$ for each degree of eye movement. In this preliminary phase, the exact positions of the presented objects were not relevant. Our algorithm was focused on identifying parameters of variations in the recorded signals, such as amplitude and width for each saccade as described in previous section.

For the current paradigm, the training phase consists of next steps: 10 trials for each left, right, top and bottom eye movement events and 10 trials for strong double blinking event. These events were executed in the following order: left, right, top, bottom and double blink. A strong blink has an amplitude higher than any other eye activity. A double blink event was chosen because this event is executed only when the user has the intention to execute it. For each action, 7 seconds were allocated, 2 seconds for executing the command and 5 seconds for resting (a "Wait" message was displayed on the monitor).

The testing phase consisted in presenting an arrow on the monitor for indicating an object to be selected. Objects were drawn in red and, when a trial was successful, the object was changing its colour in green in order to give a visual feedback to the user (he knew that the trial was successful) thus being stimulated to pay attention to application. If the selection trial of an object was unsuccessful, the user was instructed to continue to focus on that object in order to execute the double blink event (gripper closure). This method was chosen in order to test independently the recognition algorithms, thus calculating the accuracy rate for each command. If the closure trial was successful, the object was changing its color in blue.

3.5 EOG-EEG-Based Control Paradigm

The difference between the two proposed paradigms is only at the level of gripper closure events. For this paradigm, the training step consisted of the following actions: 10 trials for each left, right, top and bottom eye movement events and 10 trials for eye closure events. These commands were executed similarly to previous EOG-based described commands.

EEG signals are used for detecting eye closure events. One electrode was used for signal acquisition, the electrode being placed at O2 location according to the international 10-20 system [15]. EEG signals are divided into five frequency bands: alpha, beta, delta, theta and gamma. Alpha is the relevant band for our application and its range is from 8 to 12 Hz. An increase in the signal amplitude can be detected when the user has closed eyes; recognition algorithm is detailed in section 3.3. EEG signals were bandpass filtered between 8 Hz and 12 Hz; also an additional 50 Hz notch filter was enabled to suppress line noise.

The testing phase is similar to the previously described strategy; the gripper closure activation process represents the difference between the paradigms. The user was instructed to close his eyes for at least 1 second.

3.6 Results

Eight subjects (5 male and 3 female; age 23-28) took part in a series of experiments. One of the subjects had prior experience with EOG systems, but none of the other subjects had prior experience with EOG or EEG systems. All subjects had tested both paradigms in order to compare them, regarding accuracy and comfort for user. Five subjects started with first paradigm and the rest of them with the second paradigm, and in the next sessions, they tested the other paradigm. This strategy was used in order not to promote only a paradigm. Each subject tested each paradigm with a minimum distance of four days between recording sessions.

The training setup is described above in the EOG-based paradigm. For both paradigms, the testing setup is as follows: 40 left, 40 right, 40 top and 40 bottom actions and 160 closure actions (1 instruction for each selection command; a closure action is represented by a double blink or eye closure event). Selection commands were presented to user in a pseudo-randomly manner, the total count for each command was identical, but commands were presented in a random order. The timing for this setup is the following: 2 seconds for selection, 2 seconds wait time (with eyes focused on the selected object), 2 seconds for gripper closing action and finally 5 seconds of waiting time when the subject was changing his gaze on the centre of the monitor. One minute delay was inserted between the training and testing phases.

For the EOG-based paradigm, the obtained results revealed a maximum accuracy for selection commands of 86.25% (average between all selection commands from one subject), whereas the average accuracy was of 78.60% (average between all selection commands from all subjects). For double blink events, the maximum achieved accuracy was 93%, whereas the average for all subjects was 79.05%.

For the EOG-EEG-based paradigm, the maximum accuracy achieved for selection commands was of 93.75% and an average of 86.50%. For eye closure events, the maximum accuracy achieved was of 98%, whereas the average was 90.9%.

For further analysis during each test, acquired signals were stored in a separate file, and so were the values from training and testing phases. After the second session, each subject was asked to answer some questions about the tested paradigms. Relevant answers and final conclusions are stated in the next section.

4 Conclusions and Future Work

These preliminary results revealed conclusions converted in future possible improvements for presented controlling paradigms. Comparing the two proposed paradigms, we notice a major difference for selection events in terms of accuracy. This difference might appear due to the user's fatigue when using the EOG-based system. Also, the difference is seen for average values. Most of the subjects presented an increase in the selection accuracy of over 7.5% in case of the EOG-EEG paradigm. One of the subjects presented an accuracy rate increase of over 10%, two of them of over 14% and a single one presented a decrease in the accuracy rate of 3%.

Considering commands for gripper closure, we conclude there is a difference between maximum accuracy rates and also for average rates. It seems that using eye closure events (through EEG signals), subjects achieved a higher accuracy rate. From the answers given by subjects, we conclude that the EOG-EEG-based paradigm was preferred by subjects; results also confirmed this fact. They argued that this combination was less tiring.

Future work is related to refining developed algorithms, in order to increase recognition accuracy rates. Current system will evolve in an asynchronous one, allowing the user to select a desired object at will. Some new tests will be conducted in order to finally choose the best paradigm for our project, considering also the fact that many objects for selection will be added in the next applications, and also new commands will be integrated.

Acknowledgments. This paper is supported by the Sectoral Operational Programme Human Resources Development (SOP HRD), financed from the European Social Fund and by the Romanian Government under the contract number POSDRU/88/1.5/S/59321 for author (1) and by the research project IREAL, contract no. 97/2007, id: 132, funded by the Romanian Council for Research CNCSIS for author (2).

References

1. European Parliament,
http://www.europarl.europa.eu/factsheets/4_8_8_en.htm
2. Wolpaw, J.R., Birbaumer, N., McFarland, D.J., Pfurtscheller, G., Vaughan, T.M.: Brain-computer interfaces for communication and control. *J. Clin. Neurophysiol.* 113(6), 767–791 (2002)
3. Birbaumer, N.: Breaking the silence: Brain-computer interfaces (BCI) for communication and motor control. *Psychophysiology* 43, 517–532 (2006)
4. Donchin, E., Spencer, K.M., Wijesinghe, R.: The Mental Prosthesis: Assessing the Speed of a P300-Based Brain-Computer Interface. *IEEE Trans. Rehab. Eng.* 8, 174–179 (2000)

5. Blankertz, B., Dornhege, G., Krauledat, M., Schroder, M., Williamson, J., Murray-Smith, R., Muller, K.R.: The Berlin brain-computer interface presents the novel mental typewriter hex-o-spell. In: Proceedings of the 3rd International Brain-Computer Interface Workshop and Training Course, Verlag der Technischen Universitat Graz, pp. 108–109 (2006)
6. Vanacker, G., Millan, J., del, R., Lew, E., Ferrez, P.W., Galan Moles, F., Philips, J., Van Brussel, H., Nuttin, M.: Context-Based Filtering for Assisted Brainactuated Wheelchair driving. In: Computational Intelligence and Neuroscience, Hindawi Publishing Corporation (2007)
7. Dhillon, H.S., Singla, R., Rekhi, N.S., Jha, R.: EOG and EMG based virtual keyboard: A brain-computer interface. In: 2nd IEEE International Conference on Computer Science and Information Technology, pp. 259–262. IEEE Press, Los Alamitos (2001)
8. Blankertz, B., Krauledat, M., Dornhege, G., Williamson, J., Murray-Smith, R., Muller, K.R.: A Note on Brain Actuated Spelling with the Berlin Brain-Computer Interface. In: Stephanidis, C. (ed.) UAHCI 2007 (Part II). LNCS, vol. 4555, pp. 759–768. Springer, Heidelberg (2007)
9. Guger, C., Harkam, W., Hertnaes, C., Pfurtscheller, G.: Prosthetic Control by an EEG-based Brain-Computer Interface (BCI). In: Proceedings AAATE 5th European Conference for the Advancement of Assistive Technology. Dusseldorf, Germany (1999)
10. Vaughan, T.M., McFarland, D.J., Schalk, G., Sarnacki, W.A., Krusienski, D.J., Sellers, E.W., Wolpaw, J.R.: The Wadsworth BCI Research and Development Program: At Home with BCI. IEEE Trans. on Neural Systems and Rehab. Eng. 14(2), 229–233 (2006)
11. Ebrahimi, T., Vesin, J.-M., Garcia, G.: Brain-Computer Interface in Multimedia Communication. IEEE Signal Processing Magazine 20(1), 14–24 (2003)
12. Leeb, R., Scherer, R., Friedman, D., Lee, F., Keinrath, C., Bischof, H., Slater, M., Pfurtscheller, G.: Combining BCI and Virtual Reality: Scouting Virtual Worlds. In: Dornhege, G., Millan, J.d.R., Hinterberger, T., Mcfarland, D.J., Müller, K.R. (eds.) Towards Brain-Computer Interfacing. MIT Press, Cambridge (2007)
13. Barea, R., Boquete, L., Mazo, M., Lopez, E.: System for Assisted Mobility Using Eye Movements. IEEE Trans. on Neural Systems and Rehab. Eng. 10(4), 209–218 (2002)
14. Duguleana, M., Mogan, G.: Using Eye Blinking for EOG-Based Robot Control. In: Camarinha-Matos, L.M., Pereira, P., Ribeiro, L. (eds.) DoCEIS 2010. IFIP AICT, vol. 314, pp. 343–350. Springer, Heidelberg (2010)
15. Jasper, H.: Ten-twenty Electrode System of the International Federation. Electroencephalography. J. Clin. Neurophysiol. 10, 371–375 (1958)
16. XVR development environment, <http://www.vrmedia.it/Xvr.htm>
17. Guger Technologies, <http://gtec.at/>

Robot Emotional State through Bayesian Visuo-Auditory Perception

José Augusto Prado¹, Carlos Simplício^{1,2}, and Jorge Dias¹

¹ Instituto de Sistemas e Robotica ISR, FCT-UC, Universidade de Coimbra, Portugal

² Instituto Politécnico de Leiria, Portugal

{jaugusro, jorge}@isr.uc.pt, carlos.simplicio@ipleiria.pt

Abstract. In this paper we focus on auditory analysis as the sensory stimulus, and on vocalization synthesis as the output signal. Our scenario is to have one robot interacting with one human through vocalization channel. Notice that vocalization is far beyond speech; while speech analysis would give us what was said, vocalization analysis gives us how was said. A social robot shall be able to perform actions in different manners according to its emotional state. Thus we propose a novel Bayesian approach to determine the emotional state the robot shall assume according to how the interlocutor is talking to it. Results shows that the classification happens as expected converging to the correct decision after two iterations.

Keywords: Bayesian Approach, Auditory Perception, Robot Emotional State, Vocalization.

1 Introduction

In the context of human robot interaction, a core problem is how to reduce the estrangement between humans and machines. In order to do this, recently researchers are investigating how to endow the robots an emotional feedback. There has never been any doubt about the importance of emotions in human behavior, especially in human relationships. The past decade, however, has seen a great deal of progress in developing computational theories of emotion that can be applied to building robots and avatars that interact emotionally with humans. According to the main stream of such theories [1], emotions are much intertwined with other cognitive processing, both as antecedents (emotions affect cognition) and consequences (cognition affects emotions). In our scenario, a pre-defined story board exists, which the human and the robot shall follow, though removing the importance of what is said and focusing the experiments on the detection of emotion. In the simplest case, robot will mimic the detected emotion.

2 Contribution to Sustainability

Schroder et. al. [2] presented the SEMAINE API as a framework for enabling the creation of simple or complex emotion oriented systems. Their framework is rooted in

the understanding that the use of standard formats is beneficial for interoperability and reuse of components. They show how system integration and reuse of components can work in practice. An implementation of a dialogue system was done using a 2D displayed avatar and speech interface. More work is needed in order to make the SEMAINE API fully suitable for a broad range of applications in the area of emotion-aware systems [2]. Classifying emotions in human dialogs was studied by Min [3] presenting a comparison between various acoustic feature sets and classification algorithms for classifying spoken utterances based on the emotional state of the speaker. Later, Wang [4] presented an emotion recognition system to classify human emotional state from audiovisual signals. The strategy was to extract prosodic, mel-frequency Cepstral coefficient, and formant frequency features to represent the audio characteristics of the emotional speech. A face detection scheme based on HSV color model was used to detect the face from the background. The facial expressions were represented by Gabor wavelet features. This proposed emotional recognition system was tested and had an overall recognition accuracy of 82.14% of true positives. Recently, Cowie [5] it was described a multi-cue, dynamic approach to detect emotion in video sequences. Recognition was performed via a recurrent neural network.

Our approach presents a novel probabilistic model for emotion classification based on vocalization analysis and Bayesian Networks applied for Human Robot Interaction. Our prototype robot can be seen in figure 1. Furthermore, we propose a model for integration of two modalities (visual and aural), more specifically facial expression analysis following Ekman [6] and vocalization analysis.



Fig. 1. Our prototype robot

3 Emotional States

Spinozza [7], during the seventeenth century, proposed a definition of how human emotions behave. His work was recently continued and extended by Damasio [8] [9] who proposed an approach with the joint behavior of four groups of emotional states: three of them are related to the loss of some capability of communication. A fourth group, associated to success, was also considered. Each group contains the social emotion and the Emotional Competent Stimulus (ECS) for that emotion. Damasio did not define ECS for the neutral state. Here we propose the addition of a fifth group where the neutral state is. The four groups of emotional states proposed by Damasio [9], and plus the neutral state added by us, can be summarized as follow: *fear, anger, sad, happy and neutral*. According to Damasio [9], the emotional state can be influenced by what is happening with the individual's, and also to interlocutor emotional state. Taking this into account, our system is composed by analysis and synthesis (see figure 2). In the analysis part, we are determining what are the vocal expressions produced by the human. Later in the synthesis part, the emotional state is established and the reaction is synthesized. A combination with an input from human's emotional state, which is given by facial expression analysis following Ekman [6], is also proposed on the synthesis part.

4 Bayesian Modeling

The Bayesian approach is characterized by assigning probabilities to characterize the degree of belief associated with the state of the world. Bayesian approach defines how new information should be combined with prior beliefs and how information from several modalities shall be integrated. Bayesian decision theory defines how our beliefs should be combined with our objectives to make optimal decisions.

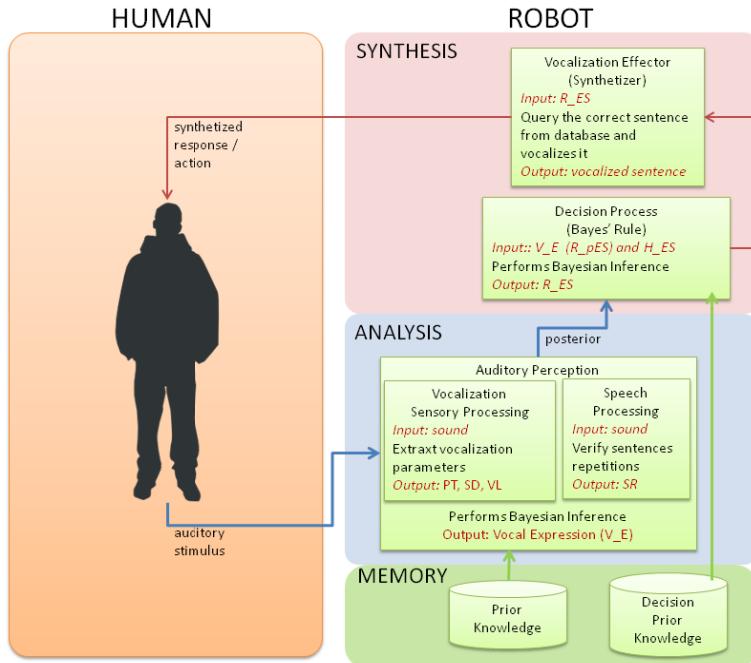


Fig. 2. Analysis and synthesis' system schema

Figure 2 shows the system implementation modules. Auditory perception feature extraction and the Bayesian model are described on section 4.1. Synthesis will be briefly presented in section 4.2.

4.1 Human Vocalization Analysis

In our model of auditory perception, the vocalization analysis classifies a vocal expression. The robot needs to be capable of classifying among the possible vocal expressions, which are in the same scope as the facial expressions as defined by Ekman [6]: {*anger, fear, happy, sad, neutral*}.

4.1.1 Vocalization Analysis

All waves are effectively combinations of sinusoids. The Fourier transform takes a waveform and turns it into a function describing which sinusoids are present in the

waveform. So, as one can see in figure 3, a digitalized sound wave comes with positive and negative values. However, it is only in the frequency of oscillation of the signal that sound can exist. Obviously, a single sample cannot represent any oscillation.

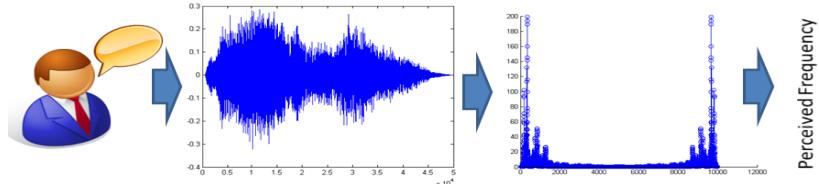


Fig. 3. Sample recorded speech waveform of an utterance, sad. The x axis is the number of samples while the y axis is the amplitude in dB. A sampling frequency of 16000Hz was used; it has 49679 samples and 3.1049 seconds of duration. At right a FFT of the interval from 1 to 2 seconds; after applying the correlation method presented by Sondhi [10] it is possible to get the perceived frequency.

In order to classify emotions from a waveform, first it is necessary to extract features from it. Here we define which features we are going to extract and also the Bayesian network to structure the relationship among them (figure 4).

There are several methods to extract pitch [10][14][15][16]: zero-crossing, autocorrelation function, cepstrum, average magnitude differential function, comb transformation, FIR filter method of periodic prediction. Lopes [11] [12] [13] extensively studied vocal tract's length normalization using pitch's features for it.

4.1.2 Auditory Perception Bayesian Network

To classify the vocal expressions performed by the human, a Bayesian network was developed. The structure of this network of two levels is illustrated in figure 4. A vocal expression will be classified after a sentence finish. In other words, for the Bayesian network, the time 1 is just after sentence 1 is completed, time 2 is just after sentence 2 is completed; and so one. This is independent of each sentence's (real time) duration. Now it is implemented over Matlab and off line, so we don't need a state switching since the sentences are recorded separated. In future, to determine the state switching, we expect to use a silence detector as presented by Hoelper [17]. If the silence period is bigger than a threshold (3 seconds), this event may trigger the state switching.

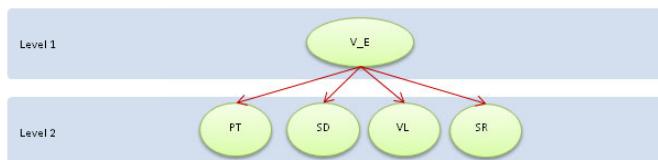


Fig. 4. Bayesian network Auditory Perception

In the Bayesian network's first level there is only one node. The global classification result obtained is provided by the belief variable associated with this node: $V_E \in \{angry, fear, happy, sad, neutral\}$; where the variable name stands from Vocal Expression. Considering the structure of the Bayesian network, the variables in the second level have as parent this one in the first level: V_E .

In the second level there are four belief variables,

- $PT \in \{short, normal, long\}$ is variable which a belief is related with Pitch. Pitch represents the perceived fundamental frequency of a sound. We are using the pitch extraction by autocorrelation method proposed by Sondhi [10].

The voice pitch changes significantly along a sentence and an important part of our voices are un-pitched, however, since the conversation follows a pre-defined story board, the mean pitch of the sentence will help to distinguish the emotional state that was there when this very sentence was spoken.

- $SD \in \{short, normal, long\}$ is variable which a belief is related with Sentence Duration. Since we know the sampling frequency ($sfreq$) of the acquired sound, and we also know the beginning and the end of each sentence, consequently the number of samples ($nsam$) then it is trivial to determine the duration in seconds by $SD = nsam/sfreq$. This variable contributes to the classification. By example, when a person speaks the same sentence with a happy emotion it usually speaks faster than with a sad emotion. For some emotional states the duration might be exactly the same, but then the other variables will contribute for the disambiguation.
- $VL \in \{low, medium, high\}$ is a belief variable which stands for Volume Level. This variable is actually the energy y of the signal, which for a continuous-time signal $x(t)$ is given by $VL = \int |x(t)|^2 dt$.
- $SR \in \{zero, one, two, three_or_more\}$ is the belief variable which stands for Sentence Repetition. It is associated with the number of sentences repetitions that the interlocutor may perform. The value of this is given by the comparison of the previous three variables along four previous times.

The following equations illustrate the joint distribution associated to the Bayesian Vocal Expressions Classifier:

$$\begin{aligned} P(V_E, PT, SD, VL, SR) &= \\ P(PT, SD, VL, SR|V_E).P(V_E) &= \\ P(PT|V_E).P(SD|V_E).P(VL|V_E).P(SR|V_E).P(V_E). \end{aligned} \tag{1}$$

The last equality can be done only if it is assumed that belief variables PT , SD , VL and SR are independent.

From the joint distribution, the posterior can be obtained by the application of the Bayes' Formula as follow:

$$\begin{aligned} P(V_E|PT, SD, VL, SR) \\ = \frac{P(PT|V_E).P(SD|V_E).P(VL|V_E).P(SR|V_E)}{P(PT, SD, VL, SR)} \end{aligned} \tag{2}$$

4.2 Robot Vocalization Synthesis

According to figure 2, the Decision Process receives as input H_ES (the Human Emotional State), which is given by the external facial expression classifier and V_E (the inferred Vocal Expression). It will take a decision according to these inputs and to the MEMORY contents.

It is assumed that the robot will initially internalize the vocal expression of the person, thus V_E implies on a robot pre-emotional state (R_pES). The Decision Process will then combine R_pES with the H_ES in order to determine R_ES (the final emotional state that the robot will assumes).

This fusion is proposed in order to determine the robotic emotional state in similar way of that established by Damasio [8] for human beings. As a consequence of the Bayesian framework, the prior knowledge will determine the balance of the fusion. This brings up to which side the decision will fall: in one side the robot is more confident in its own emotional state (from vocalization analysis); in another extreme the robot is less confident and uses the human emotional state (from facial expression analysis).

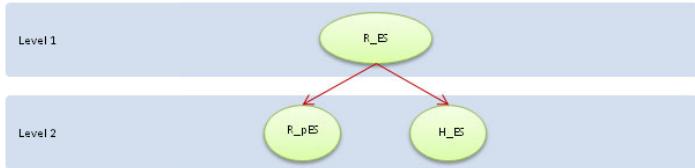


Fig. 5. Synthesis Bayesian Network

5 Results

Results of learned likelihoods for Auditory Perception.

After teaching the system, by pointing which is the correct Vocal Expression for a given input, is obtained a histogram table exactly as shown at Table 1. This histogram is the likelihood knowledge for the Bayesian algorithm to perform the inferences later: the joint distribution (see eq. 1) contains all the information needed.

Table 1. Learning for Analysis of Vocal Expressions

V_E	PT			SD			VL			SR			
	low	med	high	short	norm	long	low	med	high	0	1	2	3+
Ang	0.8	0.10	0.10	0.10	0.10	0.80	0.10	0.10	0.80	0.97	0.01	0.01	0.01
Neu	0.1	0.80	0.10	0.80	0.10	0.10	0.10	0.80	0.10	0.97	0.01	0.01	0.01
Sad	0.1	0.10	0.80	0.80	0.10	0.10	0.80	0.10	0.10	0.97	0.01	0.01	0.01
Hap	0.1	0.80	0.10	0.25	0.50	0.25	0.10	0.80	0.10	0.97	0.01	0.01	0.01
Fear	0.34	0.33	0.33	0.33	0.34	0.33	0.33	0.33	0.34	0.01	0.23	0.33	0.43

Results of Bayesian Network for Auditory Perception.

The robot is able to infer over the likelihoods (see eq. 2) when interacting to the user. The expected results for the ANALYSIS part are correct classifications of vocal

expressions according to what is expected. Convergence is also expected to appear among the time, since both are Dynamic Bayesian Networks. Figure 6 shows results of the Bayesian inference during five iterations with the following constant evidences:

Pitch=long,
VolumeLevel=low,

SentenceDuration=short,
SentenceRepetition=zero.

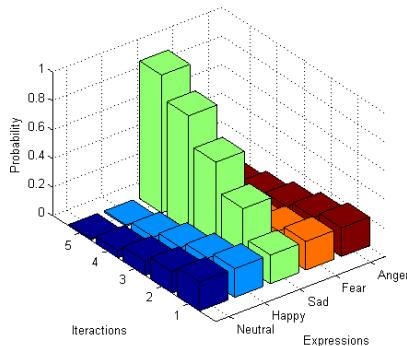


Fig. 6. Results for Classification of a Vocal Expressions (Sad) - The convergence happens after the second iteration

6 Conclusions and Future Work

This work presented a novel approach to determine robot emotional state through vocal expressions, according to philosophic references on how humans do it for themselves. The results show that correct classifications are done: the inferred emotional state is correct during an interaction between a robot and a human. This approach turns interaction more inclusive and reduces the estrangement between humans and machines. Our approach endows the robot to say the same sentence with different characteristics, according to its emotional state. We just presented one example of an utterance in sad; however, we are preparing a dataset with different sentences uttered in all the five emotional states here considered.

The current implementation of our Bayesian network is with limited values and it shows a proof of concept; however, we expect to experiment it with a larger scope of possibilities for each variable.

As the proposed model is simple, it is advantageous in particular contexts; specially multimodal fusion; where a quick (less complex) form of predicting an emotional state is better than a large model of human emotional processing.

References

1. Gratch, J., Marsella, S., Petta, P.: Modeling the cognitive antecedents and consequents of emotion. *Cognitive Systems* 10(1), 1–5 (2008)
2. Schroder, M.: The semaine api: Towards a standards-based framework for building emotion oriented systems. *Advances in Human-Computer Interaction*, article ID 319406, 21 (2010) doi:10.1155/2010/319406

3. Lee, C.M., Narayanan, S.S., Pieraccini, R.: Classifying emotions in human-machine spoken dialogs. In: ICME (2002)
4. Wang, Y., Guan, L.: Recognizing human emotion from audiovisual information. In: ICASSP. IEEE, Los Alamitos (2005)
5. Cowie, R., Douglas-Cowie, E., Karpouszis, K., Caridakis, G., Wallace, M., Kollias, S.: Recognition of Emotional States in Natural Human-Computer Interaction. Queen's University, School of Psychology (2007)
6. Ekman, P., Rosenberg, E.L.: What the face reveals: basic and applied studies of spontaneous expression using the facial action coding system (FACS), 2nd edn. Oxford University Press, Oxford (2004)
7. Spinoza, Ethics, 1677
8. Damasio, A.: Looking for Spinoza. Harcourt, Inc. (2003) ISBN 978-0-15-100557-4
9. Damasio, A.: The feeling of what happens. Harcourt, Inc., Harcourt (2000) ISBN 978-0-15-601075-7
10. Sondhi, M.M.: New methods of pitch extraction. IEEE Trans. on Audio and Electroacoustics 16(2), 262–266 (1968)
11. Lopes, C., Perdigão, F.: VTLN through frequency warping based on pitch. Revista da Sociedade Brasileira de Telecomunicações 18(1), 86–95 (2003)
12. Lopes, C., Perdigão, F.: VTLN through frequency warping based on pitch. In: Proc. IEEE International Telecommunications Symp., Natal, Brazil (September 2002)
13. Lopes, C., Perdigão, F.: On the use of pitch to perform speaker normalization. In: Proc. International Conf. on Telecommunications, Electronics and Control, Santiago de Cuba, Cuba (July 2002)
14. Zieliński, S.: Papers from work on comb transformation method of pitch detection (Description of assumptions of comb transformation Comb transformation - implementation and comparison with another pitch detection methods). Technical University of Gdansk (1997)
15. Cook, P.R., Morill, D., Smith, J.O.: An automatic pitch detection and MIDI control system for brass instruments. In: Proc. of Special Session on Automatic Pitch Detection (1992)
16. Hess, W.: Pitch Determination of Speech Signals. Springer, Heidelberg (1983)
17. Hoelper, C., Frankort, A., Erdmann, C.: Voiced/unvoiced/silence classification for offline speech coding. In: Proceedings of International Student Conference on Electrical Engineering, Prague (2003)

Manipulative Tasks Identification by Learning and Generalizing Hand Motions

Diego R. Faria, Ricardo Martins, Jorge Lobo, and Jorge Dias

Institute of Systems and Robotics, DEEC, University of Coimbra, Portugal

{diego, rmartins, jlobo, jorge}@isr.uc.pt

Abstract. In this work is proposed an approach to learn patterns and recognize a manipulative task by the extracted features among multiples observations. The diversity of information such as hand motion, fingers flexure and object trajectory are important to represent a manipulative task. By using the relevant features is possible to generate a general form of the signals that represents a specific dataset of trials. The hand motion generalization process is achieved by polynomial regression. Later, given a new observation, it is performed a classification and identification of a task by using the learned features.

Keywords: Motion Patterns, Task Recognition, Task Generalization.

1 Introduction

An important issue for modeling and recognition of human actions and behaviors are the motion patterns found during some activity. In different daily tasks the motion assumes an important key point to describe a specific action. The variety of human activity in everyday environment is very diverse; the same way that repeated performances of the same activity by the same subject can vary, similar activities performed by different individuals are also slightly different. The basic idea behind this is: if a particular motion pattern appears many times in long-term observation, this pattern must be meaningful to a user or to a task. In this work we are focusing on manipulative tasks at trajectory level to find similarities (significant patterns) given by multiples observations. The intention is learn and generalize a specific task by the hand movement including fingers motion as well as object trajectory along the task for its recognition. This application is useful for task recognition in robot imitation learning and can be applied in the future in such way that the generalized movements can be applied to other contexts by a robot. We are not going through the imitation part, but we are focusing on the ability of learning and generalization.

2 Contribution to Sustainability

During the recent years, many research fields such as human-computer interface, medical rehabilitation, robotics, surveillance, sport performance analysis have focused some of their attention to the understanding and analysis of human behaviour and

human motions. Others examples can be seen in the field of entertainment such as games that use natural user interfaces where sensors grab the human motion to interact with the game. All these research fields are searching for new solutions to improve the standard quality levels of human living conditions, generalizing the access to high quality services. Motion pattern analysis is one of the key elements to the development of those services.

The contribution of this work is an approach to learn relevant features in manipulative tasks by finding similarities among hand motions where is possible to generalize manipulative movements that can be applied to different contexts. This kind of approach can be used in the future to endow robots by imitation learning as well for recognition of a specific action to interact in a human environment to assist people in different tasks, compensating the absence of specialized human resources. Robotics can be used in medicine to assist in surgeries, rehabilitation, and also for complex task which is dangerous for human beings. These applications can contribute to the technological innovation for sustainability.

3 Related Work

In [2] is presented a programming by demonstration framework where relevant features of a given task are learned and then generalized for different contexts. Human demonstrator teaches manipulative tasks for a humanoid robot. Through GMM/BMM the signals are encoded to provide a spatio-temporal correlation. The trajectories are then generalized by using GMR. The authors in [3] presented an approach to find repeated motion patterns in long motion sequences. They state that if a point at a given instant of time, belongs to a set of repeated patterns, and then many similar shaped segments exist around that data point. They encode the characteristic point with partly locality sensitive hashing and find the repeated patterns using dynamic programming. In [4] is proposed a general approach to learn motor skills from human demonstrations. The authors have developed a library of movements by labeling each recorded movement according to task and context. By using Non-Linear differential equations they could learn and generalizing the movements.

4 Proposed Approach

Inside the neuroscience field we can find a decomposition of a typical human manipulation movement on different stages [1]. Actions phases are defined as manipulative activities involving series of primitives and events. In this work, the actions phases are used to find motion patterns in each one. In Fig.1 is possible to identify the actions phases and the events that happens among them. In each phase, it is possible to detect primitives to describe better an action. Those represented action phases are a high level segmentation of simple manipulative tasks. For the tasks that need to have in-hand manipulation (re-grasp or change the orientation of the object along the movement), the segmentation becomes more complex, the transport phase can enclose or can be changed to in-hand manipulation. Fig.2 shows the steps of the proposed approach.

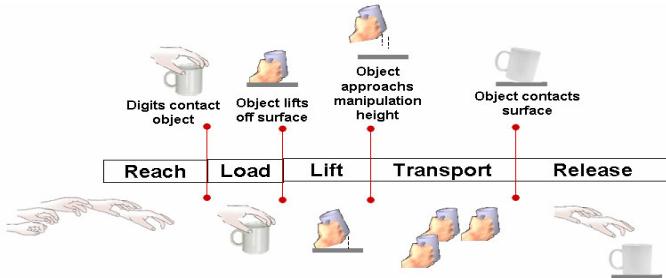


Fig. 1. Defined action phases for the manipulative tasks presented in this work

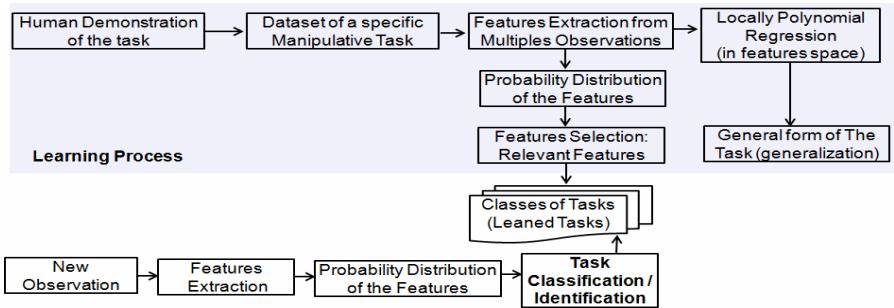
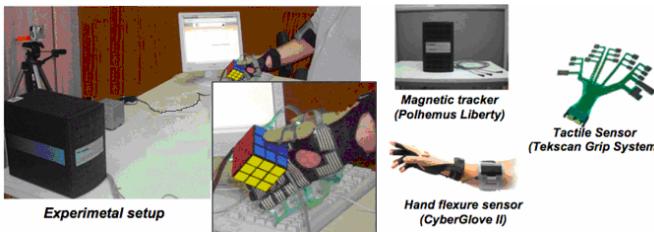


Fig. 2. Steps of the proposed approach

4.1 Experimental Setup and Data Acquisition

For the data acquisition is used: Polhemus magnetic motion tracking system [5]; TekScan grip [6] a tactile sensor for force feedback and CyberGlove II [7] for fingers flexure measurement. Each magnetic sensor has 6DoF (3D position and Euler angles). The magnetic sensors were attached to the fingertips to track the hand and fingers movements and also to track the object pose. The setup (Fig.3) for the experiments is



composed of a wooden table, and the experiments were executed by a subject seated in front of the table.

Fig. 3. Experimental Setup: Devices for the experiments

To facilitate the detection of each action phase by analyzing the sensors data, the sensors synchronization was needed. A distributed data acquisition was adopted, where a trigger defines the start and end of the acquisition. This way by looking to the

data it is possible to identify some events which enable to detect the beginning and end of an action phase in a movement. Some assumption are adopted as described in previous work [8] where some rules need to be satisfied, for example, reaching: it is defined while there is hand motion, the object is static, no force pressure from tactile sensing, and variance on the fingers flexure; load phase: hand motion, force feedback, no movement of the object.

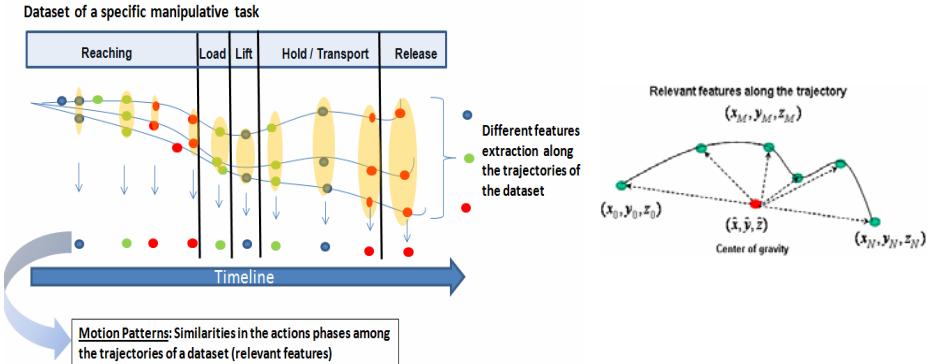


Fig. 4. – (a) Motion Patterns: Similarities detection in the action phases of the trajectories of a dataset of a manipulative task; (b) Distance of the learned features from the center of gravity

4.2 Motion Patterns: Features Extraction and Similarities among Trajectories

An example of the features selection is presented in Fig.4 (a). Given a dataset of hand trajectories of a task, we want to find the similarities among all trajectories, repeated patterns that are the relevant features to generate a generalized one. The idea is to detect features in each action phase of all trajectories of a dataset, then it is computed the probability distribution of these features. The algorithm selects the type of feature with higher occurrence in all trajectories by looking for the each feature coordinates, i.e. it is verified all first features in all trajectories and select the type of feature with high probability in that position and so on, for all second features, third until de last one. At the end, the features with high similarities among the trajectories are kept and by applying an interpolation among the features positions it is possible to have a general trajectory. The classes of features that is used to describe a trajectory are curvatures and hand orientation that vary during the task performance. In previous work [9] a probabilistic approach was developed for hand trajectory classification where curvatures and hand orientation where detected in 3D space. Here we are following the same idea for feature extraction.

4.3 Patterns from Different Sensors Modality

Other type of features is also taken into account: the fingers distances (thumb to index; index to middle and so on). The distances variance happens along to the hand trajectory, examples include hand aperture, a grasping, etc. It will also help to

differentiate each phase of a task. At each point of the fingers trajectories is computed a mean distance of the sum of squared Euclidean distances of the fingers (1). Inside of each action phase there are many 3D points so that it is possible to compute N distances. An alternative to represent each action phase is to compute the average of all computed mean distances (1).

$$D_{avg} = \frac{1}{N} \sum_{i=1}^N D_i \quad (1)$$

where

$$D = \sum_{k=1}^N (d_k)^2 \quad (2)$$

$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2} \quad (3)$$

The signals of the cyber-glove represent degrees of flexure (e.g. 0-255). An easy way to use this information is defining some grasp types such as cylindrical grasping, spherical grasping, in different levels of flexure and also defining some type of gestures, i.e. extension and flexion of the hand in different degrees of flexure (e.g. low, medium, high). After a learning stage by analyzing many observations of the same gesture it is possible to know the degree of flexure for each finger for each grasp type. In each task, it is necessary to identify the types of the defined grasping/gesture and then compute the probability distribution $P(Grasp | Observation)$ of each one along the action phases of the task.

4.4 Task Representation by General Form of the Trajectories

The representation of a dataset of a specific task at trajectory level is given by the general form of the data which is achieved after selecting the relevant features and then applying a regression on the data to generalize it. The spatio-temporal information is used to apply a polynomial regression to fit the data to have a smoothed trajectory of the manipulative task. The polynomial regression was chosen due to the curvilinear response during the fit and it can be adjusted because it is a special case of multiple linear regressions model. We are adopting the quadratic form of the model, a polynomial regression of second order. Although polynomial regression fits a nonlinear model to the data, as a statistical estimation problem, it is linear, in the sense that the regression function is linear in the unknown parameters that are estimated from the data. The general model of second order polynomial regression is given by:

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i \quad (4)$$

where $x_i = X_i - \bar{X}$ and ϵ is an unobserved random error with mean zero conditioned on a scalar variable; ϵ can be computed as error of least square fitting; β minimizes the least square error.

In our case, due to the type of trajectories, to fit correctly the curves, the regression need to be done locally, at some parts of the trajectory, Example of regression in sub-regions of the trajectories is shown in Fig.5.

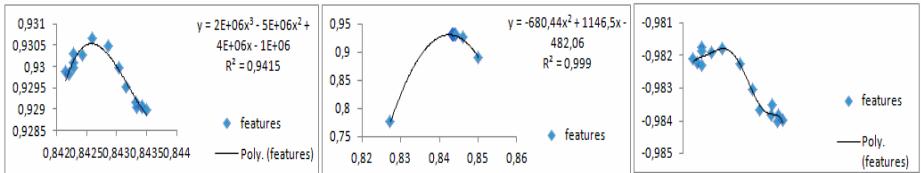


Fig. 5. – Regression applied on sub-regions of an action phase of a manipulative task. 2D view: left and middle images: x , y view; right image: x , z .

4.5 Task Identification

A new task can be recognized by matching a prototype (learned task) to a new observation or via classification where there many task classes. In the case of the hand trajectory, some properties of the learned features (translation invariance) as shown in Fig.4 (b) can be used also in the matching. The learned features is used for the matching between a prototype (generalized information) to a new observation.

Another alternative is applying continuous classification based on multiplicative updates of beliefs by Bayesian technique (5) taking in consideration the learned observations (relevant features of the general form of signals). First it is identified if the task to be classified has all action phases of the learned task, and then it is possible to classify it.

$$P(G_{k+1} | c_{k+1}, i) = \frac{P(c_{k+1} | G, i) P(c_obj_{k+1} | G, i) P(o_{k+1} | G, i) P(h_{k+1} | G, i) P(G)}{\sum_j P(g_j | c_{k+1}, c_obj_{k+1}, o_{k+1}, h_{k+1}, i)} \quad (5)$$

To understand the general classification model some definitions are done: g is a known task goal from all possible G ; c is a certain value of feature C (Curvature types) found in the hand trajectories; C_obj : curvatures found in the object trajectories; H : the grasping type learned from the data-glove signals; o : a certain value of feature O (hand orientation types) i is a given index from all possible action phases A . For more details about a methodology for some features extraction and their probability distribution see [9]. The probability $P(c | g i)$ that a feature C has certain value c can be defined by learning the probability distribution $P(C | G A)$; $P(o | g i)$ of feature O learning $P(O | G A)$; $P(h | g i)$ learning $P(H | G A)$ and $P(c_obj | g i)$ of feature C_obj learning $P(C_obj | G A)$.

5 Results

The trajectories that were used are pick-up and place and pick-up and lift. In Fig.6 (a) is shown the raw data of the used dataset correspondent to the task pick-up and place (hand trajectories); (b) shows the detected action phases using the sensors information. Fig.7 (a) shows an example of the 3D positions of the features extracted (curvatures: trajectory directions) from all observations before finding similarities; (b) relevant features selection by analyzing the probability distribution of the features to know which type of feature is more relevant, later after computing the least square among all features points of the trajectories dataset we can estimate the coordinates of them; (c) example of interpolation of the features points as a function of arc length along a space curve.

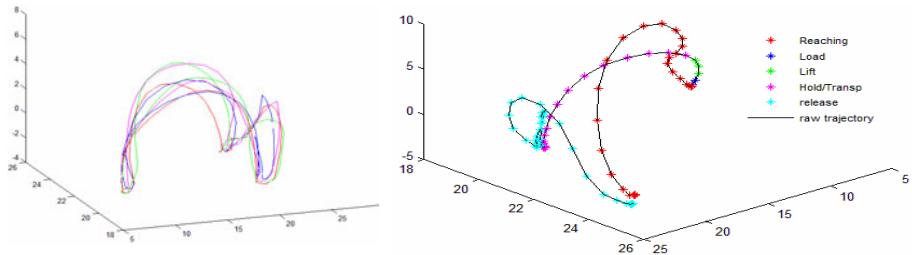


Fig. 6. – Left: Raw data(in inches): trajectories dataset (object displacement); Right: Trajectory segmentation by action phase

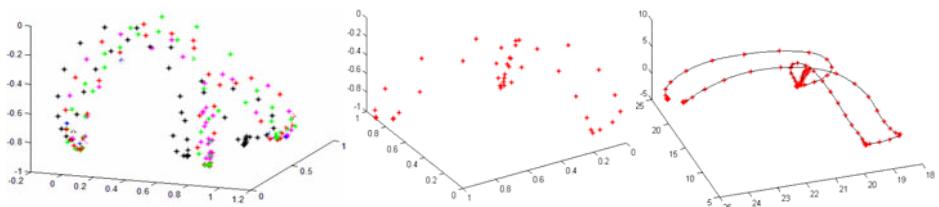


Fig. 7. – (a) Extracted Features; (b) Relevant Features (similarities among all trajectories); (c) Generalized Trajectory

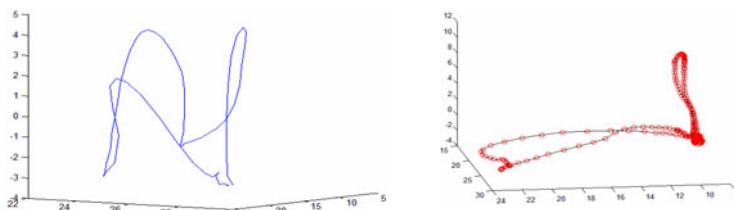


Fig. 8. – (a) New observation: trajectory to be classified (pick-up and place); (b) Trajectory of dataset pick-up and lift

Table 1. Classification Result

Action Phases	Pick-up and place %	Pick-up and lift %
Reaching	45.00	55.00
Load	48.10	51.90
Lift	59.32	40.68
Transport	69.83	30.17
Release	78.00	22.00

The 2nd and 3rd columns show the % of the new observation belonging to pick-up and place or pick-up and lift task in each instant (part of the task). We have detected the relevant features in each phase using their probabilities to classify the new observation. The new trajectory (Fig. 8(a)) is classified as pick-up and place correctly with 78%.

The actions phases for both dataset happen in different period. Given a new trajectory we want to recognize what kind of task is it. The classification variables are updated in each action phase and at the end the variables keep the final result of the classification. Table 1 shows the result of the classification of a new observation of pick-up and place. For that, just the learned curvatures features was used in the classification model (5), in the future we intend to implement the complete model to reach better results.

6 Conclusion and Future Work

In this work is proposed an approach to represent and recognize a manipulative task by multiple observations performed by a human demonstrator. By finding the relevant features of a task dataset is possible to find a general form of representing a task and also recognize it. The preliminary results motivate us to follow this proposed methodology to reach satisfactory results. In future work will be tested and evaluated the proposed approach by applying it in different trials of different manipulative tasks and different sensors signals will be used.

Acknowledgments. This work is partially supported by the European project HANDLE within the 7^o framework FP7. Diego Faria and Ricardo Martins are supported by the Portuguese Foundation for Science and Technology (FCT).

References

1. Johansson, R.S., Flanagan, J.R.: Coding and use of tactile signals from the fingertips in object manipulation tasks. *Nat. Rev. Neurosci.* 10, 345–359 (2009)
2. Calinon, S., Guenter, F., Billard, A.: On Learning, Representing and Generalizing a Task in a Humanoid Robot. *IEEE Transactions on Systems, Man and Cybernetics, Part B* 37(2), 286–298 ; Special issue on robot learning by observation
3. Ogawara, K., Tanabe, Y., Kurazume, R., Hasegawa, T.: Detecting repeated motion patterns via dynamic programming using motion density. In: Proc. 2009 IEEE Int. Conf. on Robotics and Automation (ICRA 2009), pp. 1743–1749 (2009)
4. Pastor, P., Hoffmann, H., Asfour, T., Schaal, S.: Learning and generalization of motor skills by learning from demonstration. In: Int. Conference on Robotics and Automation, ICRA 2009 (2009)
5. Polhemus Liberty 240/8 Motion Tracking System,
http://www.polhemus.com/?page=Motion_Liberty
6. TekScan Grip Sensor,
<http://www.tekscan.com/medical/system-grip.html>
7. CyberGlove II, <http://www.cyberglovesystems.com>
8. Faria, D.R., Martins, R., Dias, J.: Learning Motion Patterns from Multiple Observations along the Actions Phases of Manipulative Tasks. In: Workshop on Grasping Planning and Task Learning by Imitation: IEEE/RSJ IROS 2010, Taipei, Taiwan (October 2010)
9. Faria, D.R., Dias, J.: 3D Hand Trajectory Segmentation by Curvatures and Hand Orientation for Classification through a Probabilistic Approach. In: Proceedings of The IEEE/RSJ Int.Conf. on Intelligent Robots and Systems, IROS 2009, St. Louis, MO, USA (2009)

Evaluation of the Average Selection Speed Ratio between an Eye Tracking and a Head Tracking Interaction Interface

Florin Bărbuceanu¹, Mihai Duguleana¹, Stoianovici Vlad², and Adrian Nedelcu²

¹ Transilvania University of Brasov, Product Design and Robotics Department

² Transilvania University of Brasov, Electronics and Computers Department

{florin.barbuceanu,mihai.duguleana,stoianovici.vlad,
adrian.nedelcu}@unitbv.ro

Abstract. For a natural interaction, people immersed within a virtual environment (like a CAVE system) use multimodal input devices (i.e. pointing devices, haptic devices, 3D mouse, infrared markers and so on). In the case of physically impaired people who are limited in their ability of moving their hands, it is necessary to use other special input devices in order to be able to perform a natural interaction. For the inference of their preference or interests regarding the surrounding environment, it is possible to take in consideration the movements of their eyes or head. Based on the analysis of eye movements, an assistive high level eye tracking interface can be designed to find the intentions of the users. A natural interaction can also be performed at some extent using head movements. This work is a compared study regarding the promptness of selection between two interaction interfaces, one based on head tracking and the other based on eye tracking. Several experiments have been conducted in order to obtain a selection speed ratio during the process of selecting virtual objects. This parameter is useful in the evaluation of promptness or ergonomics of a certain selection method, provided that eyes focus almost instantly on the objects of interest, long before a selection is completed with any other kind of interaction device (i.e. mouse, pointing wand, infrared markers). For the tests, the tracking of eyes and head movements has been performed with a high speed and highly accurate head mounted eye tracker and a 6 DoF magnetic sensor attached to the head. Direction of gaze is considered with respect to the orientation of head, thus users are free to turn around or move freely during the experiments. The interaction interface based on eye tracking allows the users to make selections just by gazing at objects, while the head tracking method forces the users to turn their heads towards the objects they want to be selected.

Keywords: eye tracking, head tracking, interaction metaphor, interaction interface, virtual reality, virtual environment, CAVE system.

1 Introduction

As vision is one of the most important communication channels, vast research has been conducted lately in the area of eye tracking. Sayings which date from the early

ages state that eyes are the window towards mind. It is sometimes facile for relatives, friends or even strangers to guess someone's intentions just by looking at their eyes. Although they are input sensory channels, when they behave according to known gestures it is possible for some meaningful information to be transmitted through. This information is very useful in the implementation of assistive interaction interfaces, especially in the case of severely disabled people [1]. Attentive user interfaces (AUIs) take this information into consideration to infer user's intensions and preferences [2]. For the purpose of an increased degree of self-sufficiency in carrying out daily life activities [3], some heterogeneous environments like inhabited rooms can be controlled by disabled people through communication interfaces based on eye-tracking [4]. The users gradually shift their gaze towards a certain target until their preference is established [5]. The "cascade effect" discovered in 2003 [6] relate the gradual gaze shifts with the interest of users towards solving a given task. Due to this association between saccadic eye movements and interest, a quick determination of user's gaze at any time is essential for a consistent inference of user's interests [7]. The most important advantage of an object selection interface based on eye tracking is the promptness. Based on the fast analysis of eye movements, an attentive interface can make associations between sequentially gazed targets, time spent on each target or the path followed with the gaze, to guess the interest of the users or to determine if they are in a situation of uncertainty [2]. The *iTourist* system is able to analyze user's gaze very quickly and make associations between fixations points over the surface of an electronic map, providing a dynamic flow of information about what he/she appears to be interested in. It reacts as a very attentive humanlike guide, paying attention at all times to what the tourist is looking at [8].

2 Contribution to Sustainability

One of the most inconvenient aspects of an object selection method based on eye tracking is the imprecision of gaze estimation. Gaze position accuracy of the eye tracking system used in our experiments is between 0.5° and 1° . This means that objects located far from the user are more likely to be missed by the estimated direction of gaze, especially if they are small.

If head position and orientation can be tracked, an estimation of the user's gaze can be considered along the orientation of head. Object selection can be performed very precisely in this case if a visual feedback of head orientation is presented to the user. Provided that head movements can be successfully used as an interaction method in a virtual environment, we have conducted a set of experiments to compare the promptness of a head tracking interaction metaphor with respect to a fast eye tracking interaction metaphor. However, the head tracking method is not as fast as the one based on gaze tracking and the time delay between selections made through the two interaction interfaces is rather intuitive. The purpose of this paper is to discuss the results obtained in a series of experiments, regarding the selection speed ratio between the two interaction interfaces mentioned. It can be an instrument in the evaluation of the stress exerted on the user when using head movements to make selections. Any unnecessary load within the interaction metaphor can lead in time to fatigue, especially if the interaction metaphor is complex. Head movements are complex and require more spatial coordination, so in this case it is essential to know the amount of time users spend on the selection procedure.

3 Design of Experiments

Calibration of the eye tracker used in our experiments, ASL H6-HS-BN 6000 Eye-Track model, an accurate and high speed head mounted system, is typically made on a normal desktop screen (Fig. 1), by sequentially gazing at each one of the green points.

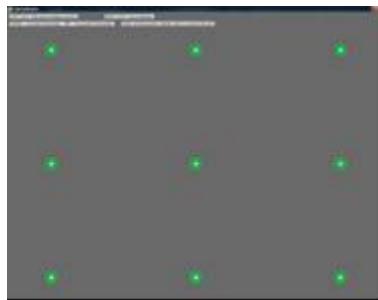


Fig. 1. Eye tracker calibration points on a desktop PC

Our experiments have been conducted on a large projection screen, normally used for visualization of stereoscopic scenes. The nine calibration points were displayed in a similar fashion as on the small desktop screen (Fig. 2), covering the visual field of the user. During the experiments, a black background was chosen for the projection screen in order not to distract the subjects in any way from the task they were assigned to (Fig. 3). On the black background a green square is displayed sequentially in 9 locations on the screen in a random fashion, so that subjects can't anticipate the next location where it will be displayed. Since the accuracy of the selection is not the subject of these experiments, the locations on the screen of the 9 points were chosen at 0.5 m one from another in order to enable a facile discrimination of each gazed objects.



Fig. 2. Eye tracking calibration points displayed on the powerwall



Fig. 3. The green square displayed on the powerwall

The scenario of the tests was very simple; the users had to perform as many selections as possible within 90 seconds. Gaze direction is represented by a bounding box, starting from user's head towards the screen, long and thick enough to collide with the green square. The system detects when the gaze direction of users falls over the green square, by testing the collision between the bounding bar of gaze direction and the green square. In Fig. 4 the frame of the bounding bar is displayed and one could notice that it intersects one of the objects in the virtual environment. In this case the bounding box of the selected object is also drawn to confirm the success of the selection.

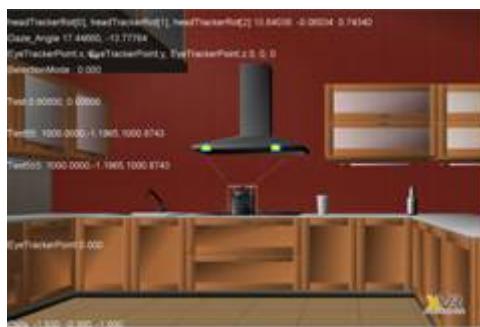


Fig. 4. The frames of the gaze direction bounding box

The bounding box of the gaze direction has one of its faces at the corresponding location of subject's head while the orientation is given either by head orientation either by gaze direction, depending on the selection method currently used. Because of this spatial disposition, some users mentioned that the bounding box can be regarded as an extension in the virtual reality of the human body, or as a self-centered pointing device. Head position and orientation are retrieved in real time by a magnetic tracker. A 6 DoF sensor is attached to the helmet (Fig. 5), thus providing a complete freedom of movement to the subjects during the experiments. The software used for visualization is XVR Studio (EXtreme Virtual Reality) developed by the VRMedia Spin Off of Scuola Superiore Sant' Anna, Italy. This software architecture has the



Fig. 5. Magnetic sensor attached to the ASL eye tracker's helmet

capability of extending its features through external dynamic link libraries (dll). External data can also be injected using socket communication. For head tracking we have used an external connection to a dynamic library written in c++ which retrieves the position and orientation of the head from a magnetic tracker. Data from the eye tracking device was transferred through an UDP port.

4 Discussion of Results

During the experiments, when a selection occurs, the counter for the number of completed selections is increased. This variable is saved in a database, along with time when each selection occurs (minute, seconds and milliseconds). The overall results obtained for each of the 10 subjects are available in Fig. 6. They clearly indicate that object selection speed ratio is considerable superior for the interaction interface based on eye tracking (Fig. 6(a)). In average, this ratio is 2.47 (Fig. 6(b)) – the blue bar). The standard deviation of the number of selections completed is 13.6 for the head tracking method and 15.7 for the method based on eye tracking.

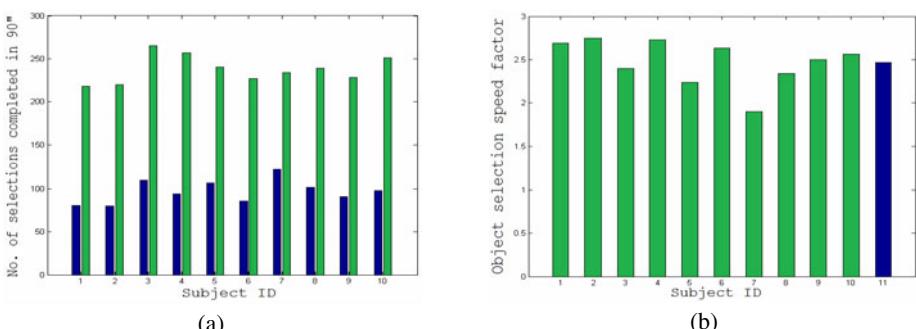


Fig. 6. Number of selections completed by each subject within 90" (a): green bars – selections made by gaze tracking; blue bars – selections made by head tracking; Selection speed ratio (b): - green bars – selection speed ratio for each subject; blue bar – average selection speed ratio

When designing a natural interaction interface it is essential to have in mind the easiness of the selection procedure, simplification, promptness and user abilities. Our contribution lies in the evaluation of the promptness of a head tracking selection interface, relative to a fast gaze tracking interface. Provided that selections of objects made with an interface based on gaze tracking are faster than any other selection method, values obtained in these experiments can be used as a point of reference for evaluation and comparison of other selection methods, in terms of selection promptness or stress exerted on the user. In the case of the head tracking selection interface, an average 2.47 ratio can be considered as high, because head movements requires complex spatial coordination, forcing the user to be more focused. This delay in combination with the constraint of wearing a sensor on the head, in time can lead to fatigue and discomfort.

Acknowledgments. This paper is supported by the Sectoral Operational Programme Human Resources Development (SOP HRD), financed from the European Social Fund and by the Romanian Government under the contract number POSDRU/6/1.5/S/6. Also special thanks to PhD. Eng. Franco Tecchia and Prof. PhD. Eng Marcello Carrozzino from the PERCRO laboratory, Pisa, Italy, for their support during the preparation of the experiments.

References

1. Barea, R., Boquete, L., Mazo, M., López, M.: Wheelchair Guidance Strategies Using EOG. *Journal of Intelligent and Robotic Systems* 34, 279–299 (2002)
2. Prendinger, H., Hyrskykari, A., et al.: Attentive interfaces for users with disabilities: eye gaze for intention and uncertainty estimation. *Univ. Access Inf. Soc.* 8, 339–354 (2009)
3. Martens, C., Prenzel, O., Gräser, A.: The Rehabilitation Robots FRIEND-I & II: Daily Life Independency through Semi-Autonomous Task-Execution. In: *Rehabilitation Robotics*. I-Tech Education Publishing, Vienna (2007)
4. Corno, F., Gale, A., Majaranta, P., Räihä, K.J.: Eye based direct interaction for environmental control in heterogeneous smart environments. *Handbook of Ambient Intelligence and Smart Environments* 9, 1117–1138 (2010)
5. Bee, N., Prendinger, H., Nakasone, A., André, E., Ishizuka, M.: AutoSelect: What You Want Is What You Get: Real-Time Processing of Visual Attention and Affect. In: André, E., Dybkjær, L., Minker, W., Neumann, H., Weber, M. (eds.) *PIT 2006. LNCS (LNAI)*, vol. 4021, pp. 40–52. Springer, Heidelberg (2006)
6. Shimojo, S., Simion, C., Shimojo, E.: Gaze bias both reflects and influences preference. *Nature Neuroscience* 6, 1317–1322 (2003)
7. Vertegaal, R., Shell, J.S., Chen, D., Mamuji, A.: Designing for augmented attention: Towards a framework for attentive user interfaces. *Computers in Human Behavior* 22, 771–789 (2006)
8. Qvarfordt, P., Zhai, S.: Conversing with the user based on eyegaze patterns. In: *Proceedings of the ACM CHI 2005 Conference on Human Factors in Computing Systems*, pp. 221–230. ACM, New York (2005)

LMA-Based Human Behaviour Analysis Using HMM

Kamrad Khoshhal¹, Hadi Aliakbarpour¹, Kamel Mekhnacha², Julien Ros²,
Joao Quintas¹, and Jorge Dias¹

¹ Institute of Systems and Robotics – University of Coimbra – Portugal

{kamrad, hadi, jquintas, jorge}@isr.uc.pt

² Probayes SAS - Montbonnot - France

{kamel.mekhnacha, julien.ros}@probayes.com

Abstract. In this paper a new body motion-based Human Behaviour Analysing (HBA) approach is proposed for the sake of events classification. Here, the interesting events are as normal and abnormal behaviours in a Automated Teller Machine (ATM) scenario. The concept of Laban Movement Analysis (LMA), which is a known human movement analysing system, is used in order to define and extract sufficient features. A two-phase probabilistic approach have been applied to model the system's state. Firstly, a Bayesian network is used to estimate LMA-based human movement parameters. Then the sequence of the obtained LMA parameters are used as the inputs of the second phase. As the second phase, the Hidden Markov Model (HMM), which is a well-known approach to deal with the time-sequential data, is used regarding the context of the ATM scenario. The achieved results prove the eligibility and efficiency of the proposed method for the surveillance applications.

Keywords: Human Behaviour Analysing, Laban Movement Analysis, HMM and Bayesian Network.

1 Introduction

HBA is demanding for many applications such surveillance systems and home-cares. Human Movement Analysis (HMA) consider as a prerequisite for the body motion-based HBA. Having human movement's properties makes it possible to interpret human behaviours, which is a more complex task. As Bobick [4] believes, human behaviour comes from a sequence of performed human motions inside a scene. It means that previous knowledge of human motion is needed, to be able to understand human behaviours in the contex of a particular scenario. Thus it seems that there is a couple of issues, namely a sequence of movement and environment parameters that need to be investigated in order to estimate the human behaviour.

Hidden Markov Model (HMM), which is kind of Dynamic Bayesian Network (DBN) methods, is a well-known method for the purpose of analysing in such a sequential movement data and it can deal with previous knowledge dependencies. Remagnino and Jones [14] used a HMM approach to model parking lot environment behaviours. Oliver et al. in [9] defined some sequences of human motions to estimate the people behaviours.

There are many elements or parameters, which can affect human behaviour in different situations. Pentland and Liu in [11] discussed how to model the human behaviour in a driving situation. They believed that it is useful to have some dynamic models for each kind of driving such as relaxed driving, tight driving, etc. then classify the driver's behaviour by comparing it with the models. Nascimento et al. in [8] described a method for recognizing some human activities in a shopping space (e.g. entering, exiting, passing and browsing). They used human motion patterns, which were achieved from a sequence of displacements of each human's blob center.

Ryoo and Aggarwal have used the HMM for different level of human behaviour understanding: primitive and complex in 2D-base space [16], [17]. A deep contribution in the field of human-machine interaction (HMI), based on the concept of LMA, is performed by Rett & Dias in [15]. In their work a Bayesian model is defined for learning and classification. The LMA is presented as a concept to identify useful features of human movements to classify human gestures.

In this paper, by getting inspiration from the previous work of Rett and Dias in [15], the concept of LMA is used in order to define and extract sufficient features. A two-phase probabilistic approach has been applied to model the system's state (see Fig. 1). Firstly, a Bayesian network is used to estimate LMA-based human movement parameters. Then a sequence of the obtained LMA parameters is used as the inputs of the second phase. As the second phase, the HMM, which is a well-known approach to deal with the time-sequential data, is used regarding the context of the ATM scenario. The achieved results prove the eligibility and efficiency of the proposed method for the surveillance applications.

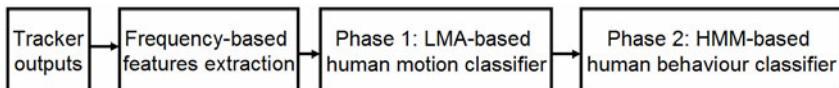


Fig. 1. The methodology diagram

This paper is arranged as following. Section 2 describes the contribution to sustainability of the paper. Section 3 presents a frequency-based feature extraction method. Then LMA concept and the human motion understanding approach are described in section 4. Section 5 describes the second phase of our HMM-based classification part, which is defined for human behavior understanding. Section 6 presents experimental part, and Section 7 closes with a conclusion and an outlook for future works.

2 Contribution to Sustainability

In this paper, a new body motion-based HBA approach is proposed for the sake of events classification. The key is that a couple of classifier was used based on frequency-domain features and LMA concept. The impact of this paper will be a reliable behaviour Analysing system using human



Fig. 2. Robberies state in an ATM scenario - PROMETHEUS dataset

body motion features to analyse different human-being events. The system will be useful in many applications especially in surveillance systems and smart-homes, which are growing very fast in the world. In this paper, the interesting events are categorized as normal and abnormal behaviours in ATM scenarios (see Fig. 2).

3 Frequency-Based Feature Extraction

We believe that the frequency-based features are suitable features to achieve the LMA.*Effort* parameters and recognize human motion, as can be seen in our previous

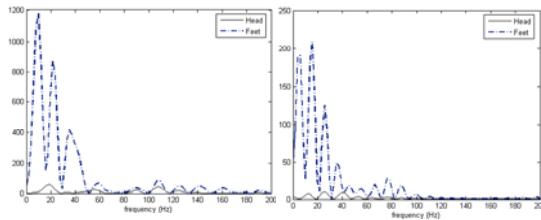


Fig. 3. PS of acceleration signals of two body parts (head and feet) for (left) running and (right) walking movement

work [6] in detail (LMA.*Effort* parameters shall be introduced in the next section). Frequency-based features can be obtained by using Fast Fourier Transform (FFT) and Power Spectrum (PS) techniques on the input signals ([13] and [5]). The acceleration signals of the body parts are supposed to be available as the input data, and then FFT and PS of those signals are extracted as can be seen in Fig. 3.

By having all these PS signals for each selected body part acceleration signal in different actions, and collecting some coefficients (peak) of the extracted signals, like [6] which collected first four coefficients of each subdomain of PS signals, we have sufficient features in our application. Thus, four features for two parts of body are defined to be used for classification of various actions. $\text{Max}_{\text{Acc } f_i \text{ pb}}^{\{ \cdot \}}$ denotes the maximum content of each i subdomain-frequency of acceleration signal for each parts of body (pb). The set of pb and subdomain frequency are defined as {Head, feet} and {(0-10), (11-20), (21-30), (31-40)} in Hz unit.

4 LMA-Based Human Motion Modelling

Laban Movement Analysis (LMA) is a known method for observing, describing, notating, and interpreting human movement, that was developed by Rudolf Laban, who is widely regarded as a pioneer of European modern dance and theorist of movement education [19] about 60 years ago. Norman Badler's group was the first group who attempted to re-formulate Laban framework in computational models since 1993 [1], [19]. Recently Dias's group also had several interesting works around LMA since 2007 [15].

The theory of LMA consists of several major components, though the available literature is not in unison about their total number. The works of Norman Badlers group [19] mentioned five major components; *Body*, *Effort*, *Space*, *Shape* and *Relationship*.

One of the most important components of LMA is *Effort* that we tried to obtain it and then based on that, reach to human behaviour. *Effort* or dynamics is a system for understanding the more subtle characteristics about the way a movement is done with respect to inner intention. The difference between punching someone in anger and reaching for a glass is slight in terms of body organization - both rely on extension of the arm. The attention to the strength of the movement, the control of the movement and the timing of the movement are very different. *Effort* has four subcategories; Time, Space, Weight and Flow, and each of them has two opposite polarities which are sudden/sustained, direct/indirect, strong/light and bounded/free, respectively (for more information: [6]).

In the first phase, LMA parameters were obtained to analyse some interesting human movements depend on the scenario and tracker outputs. Based on the collected data and our tracking outputs, we could just rely on position of feet and head. Thus, we could have just acceleration signal of these body parts.

The interesting activities in ATM scenario are standing, walking, running and falling down. In our previous work [6] based on the results, we realized that by having just a couple of body parts positions instead of six parts of body, and having just a couple of states for *Effort.time* (sudden and sustained) are not enough to distinguish the interesting activities. Thus we discretized *Effort.time* to four states, as can be seen in Fig. 4. Then the outputs of this Bayesian net are all the probabilities of the interesting activities.

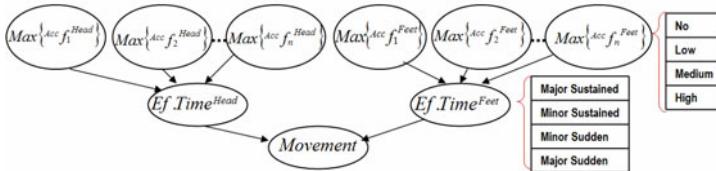


Fig. 4. LMA-based Bayesian Net

5 Concurrent HMM-Based Human Behaviour Modeling

For behaviour recognition, we are interested in detecting the current behaviour amongst N known behaviours (i.e. the behaviour library). For this purpose, using a concurrent HMM architecture is proposed.

5.1 Principle

A concurrent HMM is composed of several HMMs, each one describing one class (see Fig. 5 left). To summarize, the concurrent HMM centralizes the on-line update of the behaviour belief and contains:

1. The set of HMMs representing basic behaviour library (one HMM per behaviour);
2. The transition between behaviours model that could be either defined by hand (by an expert), or learnt from annotated data.

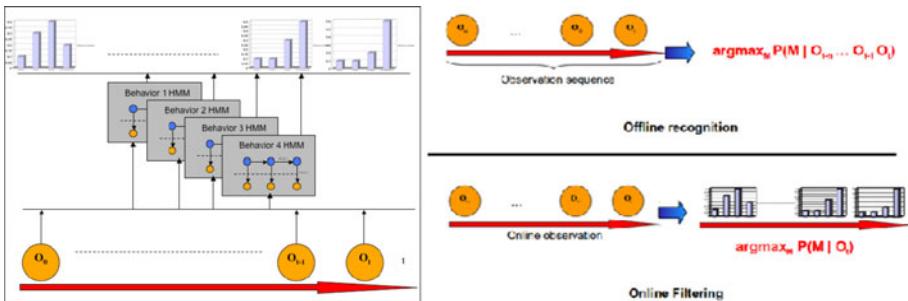


Fig. 5. Concurrent HMMs (left) Recognition Modes (right)

HMMs are used to characterize an underlying Markov chain which generates a sequence of states. The term Hidden in the HMM name comes from the fact that the sequence of states is not directly observable. Instead the states generate an observable sequence. Thus, the output depends on the current state and on previous outputs. These tools are widely used in the field of sound processing [12], gene finding and alignment in DNA sequences [10]. They were introduced by Andre Markov in [7] and developed in [2].

HMM are widely employed in the field of computer vision to recognize gesture or human behaviour [3] in these applications, the observation variables are features extracted for video data. The principle of an HMM is presented on inside of the Fig. 5 (left) (the four HMMs) in which the top (blue) circles (S) represent the state variables and the bottom circles (O) represent the observation ones in a sequence of times. In our application, each HMM describes a human behaviour and is learned using a training dataset composed of labeled observation sequences that are low-feature extracted from the LMA.

5.2 Construction/Learning

Constructing a concurrent HMM consists in:

- Learning the set of HMM models representing the behaviour library (one HMM per behaviour) using an annotated data set.
- Defining the transition matrix between the behaviours. This transition model could be either defined by hand (by an expert), or learnt from an annotated data set.

Learning the behaviour transition model is straightforward and consists in computing simple statistics (histograms) of transitions using the annotated data set. Learning the underlying HMM models (a HMM per behaviour) is more complex. It can be divided into two sub-problems:

1. Finding the optimal number of states N . The optimal number of internal states within the HMMs could be chosen by hand thanks to an expert. In this case no algorithm is needed and the learning of the HMM is reduced to the learning of its parameters. However, since an HMM is a Bayesian Network, a score that allows a compromise between fitting learning examples (D) and the ability of generalization (see the Occam Razor Principle) can be employed to find it automatically [3]. For example, the classical Bayesian Information Criterion [18] that maximizes the likelihood of the data while penalizing large size model can be used:

$$BIC(n, D) = \log(\text{likelihood}(D, n)) - \frac{1}{2} \times n\text{params}(n) \times \log(|D|)$$

In this case, the optimal number of states is given by: $n^* = \arg \max_n BIC(n, D)$

2. Learning the parameters of the HMM given N (i.e., the transition matrix $P(S_t | S_{t-1})$, the observation distribution $P(O_t | S_t)$, and the initial state distribution $P(S_0)$). The idea is find the parameters that maximize the data likelihood. For this purpose the methods generally employed are the classical EM algorithm (aka Baum-Welch algorithm in the HMM context), or the Iterative Viterbi algorithm.

5.3 Recognition

As previously emphasized, the concurrent HMM is used to recognize on-line or off-line the current behaviours amongst N known behaviours (see Fig. 5 (right)). This is easily performed by finding the HMM M that maximizes $P(M | O_{t-n}, \dots, O_t)$ for the off-line case (or $P(M | O_t)$ for the on-line case).

6 Experimental Results

In our experiments, the dataset of PROMETHEUS project has been used. Among the different various surveillance-related scenarios which exist in the database, some ATM scenarios have been selected for our intention. There is a network of cameras which observe the scene. In the ATM scenario, there are several behaviours or states which come from the involved people in the scene, such as waiting, taking money, exiting, entering and robbery at ATM area. The most interesting state in this kind of scenario usually is robbery which is an abnormal situation that can happen easily, because usually there is no any support for its security around most of the ATMs. The robbery event consists some human activities which can be defined as “when person A walks toward a person B standing close to the machine, stand by a very short time and then escape”.

Using the available tracking outputs in the database, the frequency-based features are extracted by applying the proposed method in Section 3. Then the LMA-based human motion classifier, (which is a Bayesian Network [6]) is applied to the achieved frequency-based features. Then the probabilistic outputs of the LMA-based classifier are fed to the HMM as the observation data. A couple of states, namely normal and abnormal, are defined as the outputs of the HMM classifier. The abnormal state corresponds to a situation in which a robbery is happening near the ATM and the normal state corresponds to the other activities.

In this case, the main environment parameter is the position of people related to the ATM. Thus a threshold distance for selecting the persons who are around the ATM and can involve in the scenario, is defined. By having the ATM scenarios data in PROMETHEUS dataset, we defined the robbery state as usually the robber waits in ATM's area and then goes to near of a person how is taking money from the ATM and then escapes. As can be seen in the definition, this state happens by a sequence of sub-states. Thus HMM's approach is used to model this state and normal state also. The LMA output which has four probabilities for standing, walking, running and falling down action for each person is used and collected some of those data for learning and others for classification of the HMM model.

Four scenes with different durations were collected. As we mentioned before, the interesting event which is abnormal state in this kind of scenario, is robbery. In this level, a 10 second's window on the data which will be shifted 1 second along the time is defined. By 1-second shifting the defined window along the time, 148 (windows) samples will be obtained. Between these 148 samples, there are 139 normal and 8 abnormal samples which correspond to the normal and abnormal (robbery) events, respectively. It should be mentioned that, 61 samples of normal data and 4 samples of abnormal ones have been randomly selected for learning process and the others (78 samples of normal and 4 samples of abnormal) for classification process.

Table 1. Classification results

	Normal	Robbery	%
Normal	72	6	92
Robbery	0	4	100

Table 1 presents the obtained result. It shows that to detect abnormal behaviour the collected features are appropriate and the method is very reliable, however there are some false alarm which can be reduced by using more data to learn the HMM.

7 Conclusion and Future Work

In this paper a novel body motion-based HBAapproach is proposed for the sake of events classification in a ATM security scenario. The concept of LMA, which is a known human movement analysing system, is used in order to define and extract sufficient features. A two-phase probabilistic approach has been applied to model the system's state. Firstly, a Bayesian network is used to estimate LMA-based human movement parameters. For obtaining the dependancies between a sequence of human motion, HMM approach was selected for learning and classification of human behaviours. The presented results are considerable in terms of detecting all abnormal behaviour, which is very important in the security scenarios. As the future work, we intend to apply this approach to other interesting scenarios such security and smart-home scenarios. Moreover we intend to explore human-human behaviour analysis techniques based on the LMA paramateres.

Acknowledgment. This work has been supported by the European Union within the FP7 Project PROMETHEUS, www.prometheus-FP7.eu. Hadi Ali Akbarpour is supported by the FCT (Portuguese Fundation for Science and Technology).

References

1. Badler, N.I., Phillips, C.B., Webber, B.L.: Simulating Humans: Computer Graphics, Animation, and Control. Oxford Univ. Press, Oxford (1993)
2. Baum, L.E., Petrie, T., Soules, G., Weiss, N.: A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. In: The Annals of Mathematical Statistics (1970)
3. Biem, A.: A model selection criterion for classification: Application to hmm topology optimization. In: ICDAR 2003, Washington, DC, USA, p. 104. IEEE Computer Society Press, Los Alamitos (2003)
4. Bobick, A.F.: Movement, activity and action: the role of knowledge in the perception of motion. Philosophical Trans. of the Royal Society B: Biological Sciences 352(1358), 1257–1265 (1997)
5. Cheng, F., Christmas, W., Kittler, J.: Periodic human motion description for sports video databases. In: Proceedings of the 17th ICPR (2004)
6. Khoshhal, K., Aliakbarpour, H., Quintas, J., Drews, P., Dias, J.: Probabilistic LMA-based classification of human behaviour understanding using power spectrum technique. In: 13th International Conference on Information Fusion 2010, UK (July 2010)
7. Markov, A.: An example of statistical investigation of the text ‘eugene onegin’ concerning the connection of samples in chains. Lecture at the, Royal Academy of Sciences, St. Petersburg
8. Nascimento, J.C., Figueiredo, M.A.T., Marques, J.S.: Segmentation and classification of human activities. In: Int. Workshop on Human Activity Recognition and Modelling, UK, (2005)
9. Oliver, N.M., Rosario, B., Pentland, A.: A Bayesian computer vision system for modeling human interactions. IEEE Trans. on Pattern Analysis and Machine Intelligence (2000)
10. Pachter, L., Alexandersson, M., Cawley, S.: Applications of generalized pair hidden markov models to alignment and gene finding problems. In: RECOMB 2005, NY, USA, pp. 241–248 (2005)
11. Pentland, A., Liu, A.: Modeling and prediction of human behavior. IEEE Intelligent vehicles 95, 350–355 (1995)
12. Rabiner, L.R.: A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. pp. 267–296 (1990)
13. Ragheb, H., Velastin, S., Remagnino, P., Ellis, T.: Human action recognition using robust power spectrum features. In: 15th IEEE Int. Conf. on Image Processing, pp. 753–756 (2008)
14. Remagnino, P., Jones, G.A.: Classifying surveillance events from attributes and behaviour. In: Proceedings of the Biritish Machine Vision Conference, Manchester, pp. 685–694 (2001)
15. Rett, J., Dias, J., Ahuactzin, J.-M.: Laban Movement Analysis using a Bayesian model and perspective projections. Brain, Vision and AI (2008)
16. Ryoo, M.S., Aggarwal, J.K.: Recognition of composite human activities through context-free grammar based representation. In: CVPR 2006, pp. 1709–1718 (2006)
17. Ryoo, M.S., Aggarwal, J.K.: Recognition of high-level group activities based on activities of individual members. In: WMVC 2008, pp. 1–8. IEEE Computer Society Press, Los Alamitos (2008)
18. Schwarz, G.: Estimating the dimension of a model. The Annals of Statistics 6(2), 461–464 (1978)
19. Zhao, L., Badler, N.I.: Acquiring and validating motion qualities from live limb gestures. Graphical Models, 1–16 (2005)

Daily Activity Model for Ambient Assisted Living

GuoQing Yin and Dietmar Bruckner

Institute of Computer Technology, Vienna University of Technology,
Gußhausstraße 27-29,
A-1040 Vienna, Austria
{Yin,Bruckner}@ict.tuwien.ac.at

Abstract. We propose a novel way for ambient assisted living: a system that with motion detector to observe the daily activities of the elderly, build the daily activity model of the user. In case of unusual activities the system send alarm signal to caregiver. The problems with this approach to build such a model: firstly, the activities of the user are random and dynamic distributed, that means the related data is dynamically and with huge count. Secondly, the difficulty and computational burden to get character parameters of hidden Markov model with many “states”. To deal with the first problem we take advantage of an easy filter algorithm and translate the huge dynamical data to “state” data. Secondly according the limited output of distinct observation symbols per state, we reduced the work to research the observation symbol probability distribution. Furthermore the forward algorithm used to calculate the probability of observed sequence according the build model.

Keywords: Ambient assisted living, forward algorithm, hidden Markov model.

1 Introduction

The aging problem is very important for society [1]. Ambient assisted living is one of the ways to solve the problem. There are many ideas for ambient assisted living, such as: robotic and computer for elderly, video surveillance, wearing sensors for elderly. In paper [2] a conversational robot is developed in order to increase the enjoyment of the elderly in their daily living. An intelligent, dynamically facility introduced in paper [3], which helps the elderly user to browse the internet. In paper [4] a video surveillance system is proposed. It aimed to fall detection of the elderly. A ring sensor will be introduced in paper [5], it is a 24 hour tele-nursing system.

In this paper a novel way for ambient assisted living will be introduced: a system with motion detector which installed in the living environment of the elderly. The system observes the activities of the user and builds the daily activities model of the user. According the model in case of unusual activities happened the system will send alarm signal to caregiver.

Hidden Markov model will be used to build the activities model. There are many papers about hidden Markov model: paper [6] explained the basic definition of Markov chain and the hidden Markov model, furthermore the applications of HMM. The EM Algorithm and parameter estimation for hidden Markov model described in paper

[7]. In the paper [8] the author reviews the hidden Markov model and shows its application in speech recognition. A nonstationary hidden Markov model explored in paper [9], here the dynamic transition probability parameter $A(\tau) = \{a_{ij}(\tau)\}$ is a function of time duration τ . To analyze the motion detector data and learn the behavior of the user the papers [10] and [11] adopt the hidden Markov model. The authors in paper [10] take advantage of semantic symbols and build probability model in building automation systems.

2 Contribution to Sustainability and Technological Innovation

Modern technology makes Ambient Assisted Living (AAL) possible. For example in paper [4] a video surveillance system is proposed, in paper [5] a ring sensor is introduced, and similarly another type of wear sensor “alarm button on the wrist” be used to help elderly in emergency. But because of privacy issues, visual tracking is not attractive and because of the elderly have their own psychological and physiological problems, such as memory disorder (forget to wear sensor some day or forget where the sensor is), action obstacles (cannot operate an alarm device in time or in extreme situation – unconsciousness – even cannot press an alarm button on the wrist). Non-intrusive sensing is a better way to deal with these problems, such as paper [12] which a PAS (Personal Assistant System) presented. In project ATTEND (AdapTive scenario recogniTion for Emergency and Need Detection) we avoid to use camera and microphone, without sensor to be wear on the body of the user, and nothing should to be activated by the user. Just non-intrusive sensors such as motion detector or door contactor are used. Different from the paper [12] which introduced a real-time system our idea is that according the relative stable life style of elderly, we gather the activity data of user a longer time interval. Through data analyze and unsupervised learning to build the activity model of the elderly. In case of unusual situation happened the system can send alarm to care giver. This paper presents how we analyze the data from motion detector, with hidden Markov model and forward algorithm to build activity model of the user and to analyze the result.

3 Translate Raw Dynamic Data to State Data

The used motion detector installed in the living room and it works with such principle: if the motion detector detects activities of the elderly, it will send sensor value “1” to controller, others it will keep silence with value “0”. Because of the activities of the elderly are random and dynamically distributed, for example a user got up yesterday about 7 o’clock and moved around in the living room about 2 hours but today the user gets up about 7:30 and he has activities about 1 hour but discontinuous. In such situation the data sent from motion detector is random and dynamically distributed. In top of figure 1 there is an example with gathered real data about the activities of the user for one day. There are totally 2852 data points (sensor value “1” means activity and “0” means without activity by the user). It is difficult to treat all these data points as “state” to build an activities model of the user. What we interested is the activities in a time interval, for example in 15 or 30 minutes. So we can

translate the raw data to state date with predefined time interval. The advantage of such translating are: firstly, reduced the data count; secondly, make the model building don't fall into complex details.

The approach is: according the time label of the data points separate these data in different time interval, then gather all the data in each time interval, if the sum of the activities bigger than predefined threshold value, so the time interval has state value “1” else has value “0”. The predefined threshold value (T_{th} , $0 < T_{th} < 1$) indicated the sensitivity of the translating and the predefined time interval ($T_{interval}$) decided the count of the translated state.

- 1) The gathered sensor value according $T_{interval}$

$$T = \{t_1(1), t_2(0), t_3(1), t_4(0), t_5(1), t_6(0) \dots t_n(v)\} \quad (1)$$

Here t_n is the time point that the motion detector send value to controller ($n \geq 1$), v is the sensor value itself, it has value “0” or “1”.

- 2) The activities duration between sensor value

$$\Delta T = (t_n(0) - t_{n-1}(1)) \quad (2)$$

- 3) The sum of the activities duration in $T_{interval}$

$$T_{sum} = \sum(\Delta T) \quad (3)$$

- 4) Deciding if the time interval gets value “1” or “0”.

$$\text{If } T_{sum} \geq T_{th} * T_{interval} \text{ } S_{ix} = 1; \text{ If } T_{sum} < T_{th} * T_{interval} \text{ } S_{ix} = 0 \quad (4)$$

S_{ix} is the state value that the interval should take. Here “ ix ” is the interval count (index).

In middle of figure1 is the translating result with above method. The interval begin with the dotted line and end of the dash-dot line (or between dash-dot lines) indicated in the interval has state value “1” (it means activity from the user), others has state value “0” (it means without activity from the user). Here the $T_{interval}$ is 30 minutes, so

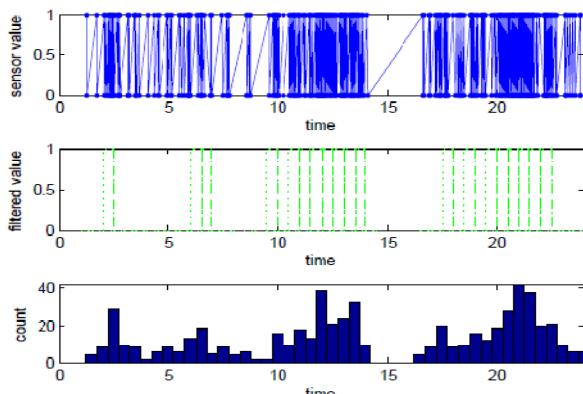


Fig. 1. Raw date, translated state data and the histogram of raw data

there are 48 states value in 24 hours. The data count reduced from 2852 data points to 48 states value. From the state value we know that the user has activities from 2 to 2:30, from 6 to 7:00, from 9:30 to 14, from 17:30 to 22:30 has activities, the other time of the day is still.

The bottom of the figure 1 is the histogram from raw data, it used to compare the translated state data with raw data.

4 Hidden Markov Model and Forward Algorithm

A hidden Markov model [8] can be characterized by following parameters.

- 1) The number of states N .
- 2) The number of output distinct observation symbols each state M .
- 3) The state transition probability distribution matrix $A = \{p_{ij}\}$.

$$p_{ij} = p \{Q_{t+1} = j | Q_t = i\}, 0 \leq p_{ij} \leq 1, \sum_{j=1}^N p_{ij} = 1, 1 \leq i, j \leq N \quad (5)$$

Here Q_t is the current state at time t. For example if $N=2$, $p_{11} = 0.4$, so $p_{12} = 0.6$.

- 4) The state emission probability distribution matrix $B = \{b_{ik}\}$.

$$b_{ik} = p \{O_t = k | Q_t = i\}, 1 \leq i \leq N, 1 \leq k \leq M \quad (6)$$

Here O_t is the output symbol at time t.

- 5) The initial state distribution $\pi = \{\pi_i\}$.

$$\pi_i = p \{Q_0 = i\} \quad (7)$$

According the parameter $\lambda (\pi, A, B)$ with forward algorithm we can find out the probability of an observed sequence $Q^{(t)} = \{q_1, q_2, \dots, q_t\}$. Here each of the q is observable state with time label.

- 6) Get the first transition probability a_1 for $t = 1$.

$$a_1(j) = \pi(j) * b_{j1} \quad (8)$$

Here j is the observation count of each observation set and $\sum \pi(j) = 1$.

- 7) For $t \geq 2$ get the transition probability $a_t(j)$

$$a_t(j) = b_{jt} * \sum_{i=1}^n (a_{t-1}(i) * p_{ij}) \quad (9)$$

- 8) For $t \leq T$ repeat (9). Here T is the length of the sequence.

5 Result and Discussion

The test data come from motion detector which observed the activities of the elderly for one week, so according (1) to (4) we get the state data $N=336$, if we predefined $T_{th} = 0.25$, $T_{interval} = 30$ minutes. Because each state has only 2 different output "0" and "1", so $M=2$. At first the states with same states value and in same time interval will be merged together. Figure 2 shows the result.

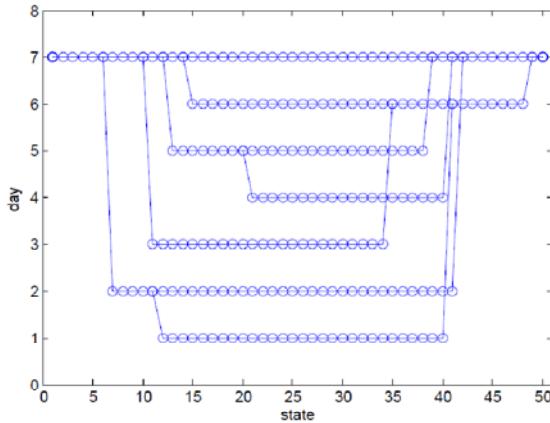


Fig. 2. The merged states in 7 days

Figure 2 shows the merging result with 336 states from 7 days. The first state and the last state on day 7 are the initial states, without states value. If the merged sequences have different value, so the sequences will be split. Different sequences will be merged again if they have same states value till to the last states. According (5) to (7) we get the parameters $\lambda = (A, B, \pi)$. These parameters present the build hidden Markov model. Here b_{ik} has the value “0” or “1” in each related state. According the build hidden Markov model with (8) and (9) the probability of an observed sequence will be find out. Figure 3 shows the result. The observation sequence is chosen from the 7 days. It compares with the sequences in the model. It is clearly the biggest probability value (logarithm value, here is -3.892) happened when the chosen day compares with itself. The smallest value is -63.81, it indicated the biggest dissimilarity between the chosen sequence and the compared sequence in the model.

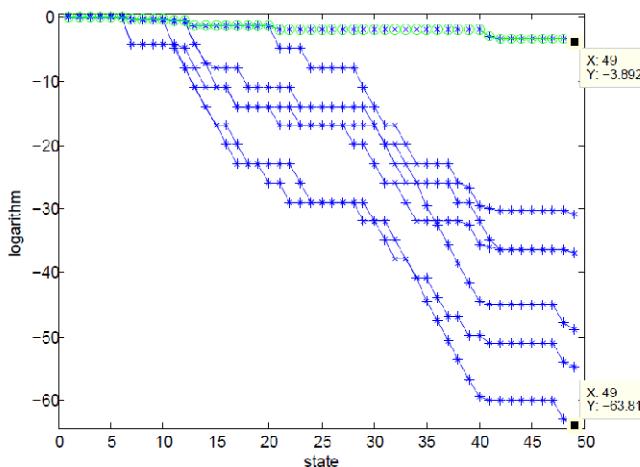


Fig. 3. The comparing result with a chosen sequence from the 7 days

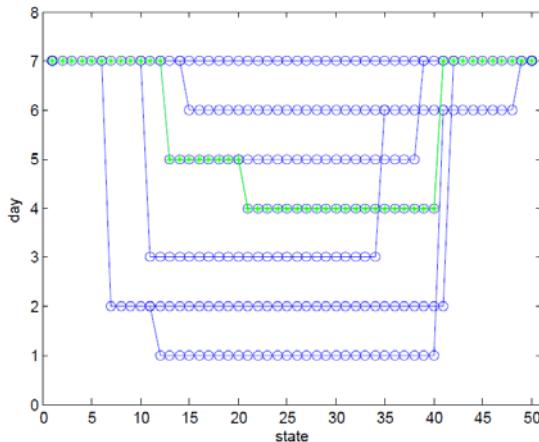


Fig. 4. The best match sequence in the build model

Figure 4 demonstrates the chosen day is the 4th day in the model according its comparing value -3.892 in figure 3.

Figure 5 illustrate the comparing result when the observed sequence is a random sequence. In such situation there must be many states didn't match to the model. That means in some time intervals the states of observed sequence have different states value as the compared sequence in the model. The parameter b_{jt} in (8) and (9) is “0” in such situation. In order to make the comparing completely (because it is perhaps after the mismatch states there are many states match again, the observed sequence and the compared sequence in the model have a high likelihood) set b_{jt} to a constant b_c . In other words if the states matched b_{jt} has value “1”; if the states mismatched set b_{jt} to a constant. So we don't need search the B matrix in (6). It must be emphasized that such simplification just used in the situation. Figure 5 displays the best match value is -57.82 and the worst match value is -84.78.

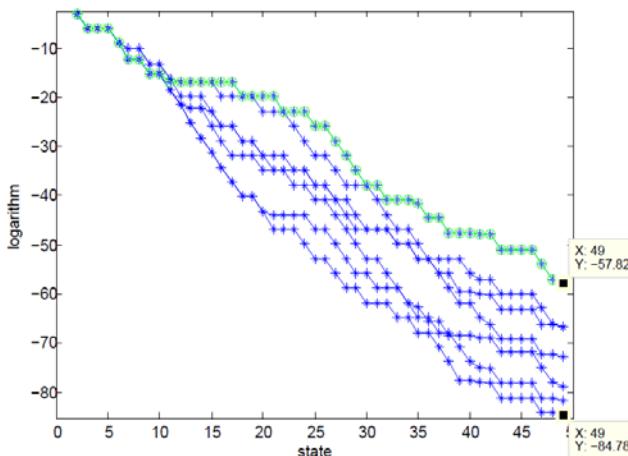


Fig. 5. The merged states in 7 days

Figure 6 shows the best match day is the 3th day in the model according its comparing value -57.82 in figure 5.

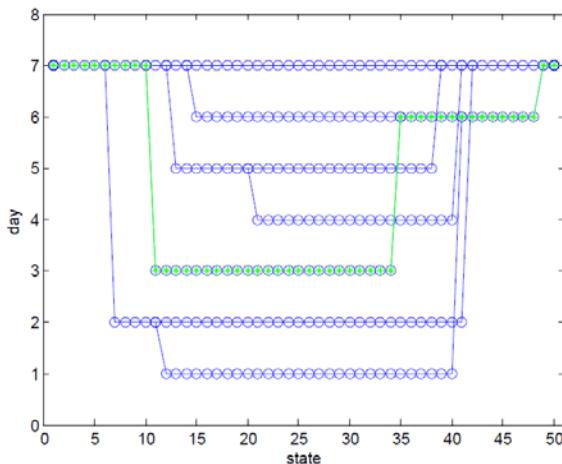


Fig. 6. The merged states in 7 days

6 Conclusion and Further Work

In this paper we propose a novel way for ambient assisted living: with the data from motion detector to build the activities model of the elderly. At first an easy filter algorithm used to translate the raw data to state data, and then get the parameters of hidden Markov model. At last the forward algorithm used to get the probability of the observation sequence comparing to the build model. In order to find out which sequence in the build model match to the observed sequence. According the special situation we reduced the work to search the B matrix in hidden Markov model but instead of a special constant b_c .

In the future the merged states from different sequences in the build model will be merged again with the consecutive states, in order to obtain a simpler but more robust model.

References

1. Burgin, M.: Age of People and Aging Problem. In: Proceedings of the 26th Annual International Conference of the IEEE EMBS, San Francisco, CA, USA, September 1- 5 (2004)
2. Heerink, M., Kroese, B., Wielinga, B., Evers, V.: Enjoyment, Intention to Use And Actual Use of a Conversational Robot by Elderly People. In: HRI 2008, Amsterdam, Netherlands, March 12-15 (2008)
3. Hunter, A., Sayers, H., McDaid, L.: An Evolvable Computer Interface for Elderly Users. In: HCI Conference on Workshop Supporting Human Memory with Interactive Systems, Lancaster, UK, September 4 (2007)

4. Foroughi, H., Aski, B.S., Pourreza, H.: Intelligent Video Surveillance for Monitoring Fall Detection of Elderly in Home Environments. In: Proceedings of 11th International Conference on Computer and Information Technology (ICCIT 2008), Khulna, Bangladesh, December 25-27 (2008)
5. Yang, B.-H., Rhee, S., Asada, H.H.: A Twenty-Four Hour Tele-Nursing System Using a Ring Sensor. In: Proc. of 1998 Int. Conf. on Robotics and Automation, Leuven, Belgium, May 16-20 (1998)
6. Bilmes, J.: What HMMs Can do, UWEE Technical Report, Number UWEETR-2002-0003 (January 2002)
7. Bilmes, J.A.: A Gentle Tutorial of the EM Algorithm and its application to Parameter Estimation for Gaussian Mixture and hidden Markov Models.In: International Computer Science Institute, Berkeley CA, 94704 (April 1998)
8. Rabiner, L.R.: A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of the IEEE 77(2) (February 1989)
9. Sin, B., Kim, J.H.: Nonstationary hidden Markov model. Signal Processing 46, 31–46 (1995)
10. Bruckner, D., Sallans, B., Russ, G.: Probability Construction of Symbols in Building Automation Systems. In: Proceedings of 2006 IEEE International Conference of Industrial Informatics INDIN 2006, Singapore, p. 6 (2006)
11. Bruckner, D., Sallans, B., Lang, R.: Behavior Learning via State Chains from Motion Detector Sensors. Bionetics (2007)
12. Hou, J.C., Wang, Q., AlShebli, B.K.: PAS: A Wireless-Enabled, Sensor-Integrated Personal Assistant System for Independent and Assisted Living. In: HCMDSS 2007, Boston, MA (June 2007)

Diagnosis in Networks of Mechatronic Agents: Validation of a Fault Propagation Model and Performance Assessment

Luis Ribeiro, José Barata, Bruno Alves, and João Ferreira

Universidade Nova de Lisboa, Faculdade de Ciências e Tecnologia, Campus da FCT-UNL,
Monte de Caparica 2829 – 516, Caparica, Portugal
`{ldr,jab}@uninova.pt, bma17241@fct.unl.pt, jpf19013@fct.unl.pt`

Abstract. Recent shop floor paradigms and approaches increasingly advocate the use of distributed systems and architectures. Plug-ability, Fault Tolerance, Robustness and Preparedness are characteristics believed to emerge by instantiation of these fundamentally new design approaches. However these features, when effectively present, often come at the cost of a greater system complexity. Enclosed in this complexity increase is a plethora of unforeseen interactions between the entities (modules) that compose the system. The purpose of this paper is, in this context, twofold: to validate a fault propagation model in random networks (that simulate the connectivity of modular shop floor systems) and assess the performance of two diagnostic approaches to expose the impact of relying in local or global information.

Keywords: Diagnosis, Complex Systems, Evolvable Systems, Agent-based Manufacturing.

1 Introduction

In recent years a considerable effort has been devoted to the research of new paradigms and development of distributed architectures to improve the shop floor response to emerging business requirements and facilitate its integration in the development of suitable and strategic-wise collaborative interactions [1-6].

At the shop floor level the main contributions can be divided in technological and paradigmatic. Industry has been the main driver for the development of platforms that explore emerging Information and Artificial Intelligence technologies (IT/AI) while the Academia has typically provided the conceptual framework that frames the usage of these platforms. Bionic Manufacturing Systems (BMS) [7], Holonic Manufacturing Systems (HMS) [8], Reconfigurable Manufacturing Systems (RMS) [9], Evolvable Assembly Systems (EAS) [10] and Evolvable Production Systems (EPS) [11] are examples of modern control approaches that envision modular systems whose components interact locally and autonomously, within their design purposes and limitations, focusing in attaining a group/social behaviour that exceeds the sum of the individual contributions. Despite the particularities, all the proposals focus in the definition of the modules' interfaces and underlying behaviours that promote a

self-organized response to production disturbances. In principle, all the main interactions are clearly defined and designed and any unforeseen behaviours are treated as an exception by a fail-safe/escape rule. The variety of systems that can be build from such building blocks is virtually unlimited. While everything seems fit in the design, the heterogeneity of the physical world (the concrete components and their working environment) rather than the paradigmatic abstraction introduce an extra level of complexity in respect to the group dynamics of the given set of components.

Traditional diagnostic tools and approaches have typically been designed to target specific systems and/or conditions. These approaches provide only a limited support for the dynamics envisioned in the paradigms detailed before. In this context, there is a lack of support tools focused both in the analysis of the group dynamics of these systems as well as in diagnosing interactions between the system components from a fault propagation and interference perspectives.

2 Technological Innovation for Sustainability

The economical growth generated with the advent of industrialization and mass production kept on feeding an insatiably society demanding low cost reliable goods.

It was believed that mass production/consumption would boost mankind to unprecedented development and sophistication. This would not be verified due to several reasons: technological advances and unscrupulous greed for profit increased unemployment and led to severe social problems; the environmental, health and safety costs were high and the progressive and general increase in customer's welfare made them increasingly demanding in respect to customized goods.

Today's society and business environment are reflexes of such changes. Being competitive is now a matter of organization sustainability. Often perceived as environmental protection, sustainable development goes beyond that and, in a broader sense, is "*development that meets the needs of the present without compromising the ability of future generations to meet their own needs*" [12]. It is a multi-dimensional and global challenge [13] were business and industry play a relevant role [14].

Mass Customization has been perceived as the excellence paradigm in industry and services it is "*the new frontier in business competition for both manufacturing and service industries. At its core is a tremendous increase in variety and customization without a corresponding increase in costs. At its limit is the mass production of individual customized goods and services. At its best, it provides strategic advantage and economic value*" [15].

Sustainable development requires a responsible implementation of such paradigm that will certainly include the adoption of innovative organization forms as detailed before. However one has first to master the intricacies of these complex systems in order to regulate them and truly profit from their dynamics maximizing, for the sake of sustainability, fundamental aspects as: equipment re-use (preventing mass disposal), uptime (preventing breakdowns and other sources of energy waste) and efficiency (producing more with less). It is, in this context, unquestionable the role of monitoring and diagnosis in what are anticipated to be the future production systems.

3 Related Literature

Targeting the area of emerging shop floor paradigms the current work gathers contributions from a multitude of areas. From a system architecture perspective the reader is referred to the articles detailed in section 1 and the references therein as well as the following development projects: SIRENA [16] – award winning project that targeted the development of a Service Infrastructure for Real time Embedded Networked Applications [17]. An Implementation of a DPWS stack [18, 19] has been produced under the framework of this project and successfully applied, at device level, in several automation test cases; SODA [20] – creation of a service oriented ecosystem based on the Devices Profile for Web Services (DPWS) framework developed under the SIRENA project; SOCRADES [21] – development of DPWS-based SOA for the next generation of industrial applications and systems; InLife [22] – development of a platform for Integrated Ambient Intelligence and Knowledge Based Services for Optimal Life-Cycle Impact of Complex Manufacturing Assembly Lines and the EUPASS project [23] focused in the study of the application of the multiagent system concept in the domain of micro assembly. The results of the EUPASS project are currently being explored under the FP7 IDEAS project that is focusing in the instantiation of the EPS paradigms in industrial controllers.

From a diagnostic point of view this research positions as complementary to the existing approaches rather than a substitute. As shall be clarified in the forthcoming section, the abstraction level considered for the purpose of performing diagnosis implies the existence of a "first line of diagnosis" that is better supported by conventional approaches. In this context, a complete review on diagnostic methods derived from the automatic control community can be found in [24] where the application of: parameter estimation, evaluation of parity relations, state estimation and principal component analysis methodologies is properly covered. A review of quantitative and qualitative history based methods where diagnosis is performed based on the previous system's faulty behaviour can be found in [25] where the application of artificial neural networks, probabilistic inference methods and expert system is discussed. Qualitative logic based diagnostic methods are covered in [26].

The fault propagation model as well as the framework for network analysis are supported by the study of complex networks [27, 28] in particular the proposed model is inspired by the work described in [29] where the conditions for cascading network effects are verified experimentally in random undirected networks.

4 Performing Diagnosis in Networks of Mechatronic Agents

When addressing the problem of performing diagnosis in the described systems from the following question arises:

Q1 which methods and tools should be developed to perform diagnosis in highly dynamic systems, like EAS/EPS, that denote physical and logical evolution and adaptation, and ensure the co-evolution/adaption of the diagnostic system without reprogramming or reconfiguring it?

Two hypothesis were set forth for the purpose of this paper:

H.1 - A self-evolving/adaptable diagnostic system for EAS/EPS can be achieved if intelligent shop floor modules explore local interaction to probabilistically infer and revise their internal states emerging at network level a consistent diagnosis.

The second hypothesis attempts to test the possibility of, rather than using local interaction and the emergent effect, use global knowledge and attempt to explain the fault propagation effect from a global network perspective.

H.2 - A self-evolving/adaptable diagnostic system for EAS/EPS can be achieved if the system composed by intelligent shop floor modules is diagnosed as an whole and the fault propagation is explained globally.

Both systems use as input the connectivity information in the network.

The system for testing hypothesis H.1 has been previously documented in [30-32] where some preliminary informal tests have been carried out to detail the significance of the nature of knowledge representation [30] and the supporting IT infrastructure [31, 32]. The system makes use of a Hidden Markov Model (HMM) [33] that from, a set of 18 possible observations, attempts to infer each agent state.

The system for testing hypothesis H.2 has been previously documented in [34] and uses a temporal logic engine to explain the most probable propagation path. Table 1 compares both approaches.

Table 1. A comparison between the systems under test

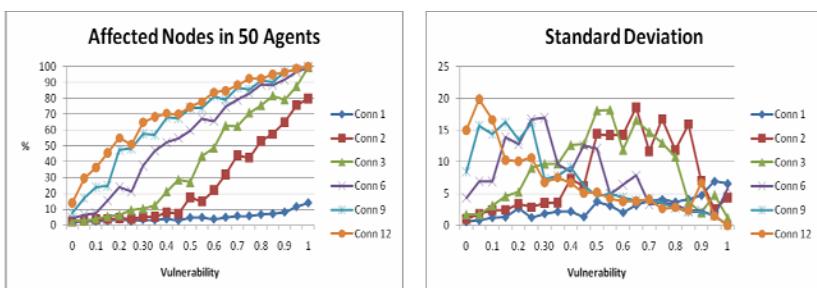
Characteristic	H.1 System	H.2 System
The reasoning process	Each agent attempts to use fault related local information (direct neighbors only) to infer its internal state. The agent can take one of five possible states that denote whether the agent is: normal (OK), suffering an uncorrelated failure (NOK), propagating a fault from which it is the origin (PFO), being affected by a fault generated elsewhere (PFOther) and a combination of the last two (PFOPFOther)	All the fault information is centrally processed. The diagnostic agent waits for the fault event to stabilize in the network and tests the best combination of system logic rules that explain the failure.
Knowledge Representation	Each agent processes the information probabilistically using an HMM that relates 18 possible observations with the five hidden states. The observations are in the form of the majority and minority of faulty neighbors both in inbound and outbound connections	The information is stored in logical statements that denote the influence of a specific type of component upon another (e.g. conveyor strongly affects another conveyor via a mechanic interaction.)

Table 1. (Continued)

Learning support	Learning is HMM is implementing by $\lambda^* = \arg \max_{\lambda} P(O \lambda)$. Learning is necessarily supervised since the a system expert has to validate the training sequences that are used to induce the model.	The system learns new logic rules as new events are detected. Each rule has a weight that denotes the frequency at which it occurs in the system. The weigh is incremented or decremented accordingly, below a certain value the rules are not used in the diagnostic process.
Complexity	The complexity of the diagnostic process is constrained at agent level and independent of the size of the network	The complexity of the diagnostic process grows with the size of the set of affected agents
Update Dynamics	Each agent performs a new diagnose asynchronously when one of its neighbors changes its state. As an whole the system takes some time to stabilize after the fault propagation stabilizes	All the information is processed at once a posteriori.

5 The Fault Propagation Model

To assess the limits of both approaches the agents abstracting the shop floor components where ran in fault simulation mode. In simulation mode one agent of the network will initiate a fault that will spread across the system. The propagation rules imply that the fault will always propagate through all possible outbound links departing from an affected agent. In practice this means that the probability that a fault affects distant nodes decays with the distance to the originating node. Some nodes in the network are more likely to be affect by the fault (vulnerable nodes) and are randomly distributed. A vulnerable node has an 80% chance of being affected as opposed to 5% chance for the remaining. Being affected by a fault does not necessarily means that it is perceived by the agent's sensor. In this context, sensor fails with a 10% chance. The percentage of vulnerable nodes was studied from 0% to



Figs. 1 and 2. The behavior of the propagation model in a network of 50 agents in respect to the percentage of vulnerable nodes (fig. 1). The standard deviation per 100 trials per vulnerability.

100% with 5% steps for random networks with the following values for average degree: 1, 2, 3, 6, 9, 12. For each value of vulnerability 100 faults (trials) were ran. The number of agents (nodes) considered was 50.

The results are depicted in Fig. 1 (the average percentage of affected nodes) and Fig. 2 (the standard deviation for each set of 100 trials). The statistical analysis of the results confirms what is intuitively anticipated in systems with such characteristics. The increase in the connectivity causes the response of the network, in respect to the number of affected nodes to shift from approximately exponential (very low values of connectivity), to approximately logarithmic. The standard deviation tends to zero for higher connectivity values as a great majority of the nodes is systematically affected. For low connectivity values the system is more sensible to the distribution of vulnerable nodes and the standard deviation grows in between 40% and 70% of vulnerable nodes. This deviation is due to the presence of clusters of vulnerable nodes in some networks.

6 Comparing Both Approaches

Both diagnostic approaches were submitted to 300 faults (trials) in networks of 50 agents to assess the performance of both systems in respect to the complexity of the network. When considering the complexity of a network several metrics have been reported in the literature [35]. Given the directed nature of the networks considered a representative measure of complexity is

$C = A/V$ (where A is the number of links in the network and V is the number of nodes).

The results of the performance assessment test are resumed in Figs. 3 and 4 where the average value for each set of trials is identified as "H.x System" where x is the number of the hypothesis under test and the confidence interval is bounded by "Lower Interval H.x" and "Upper Interval H.x" for a degree of confidence of 95%.

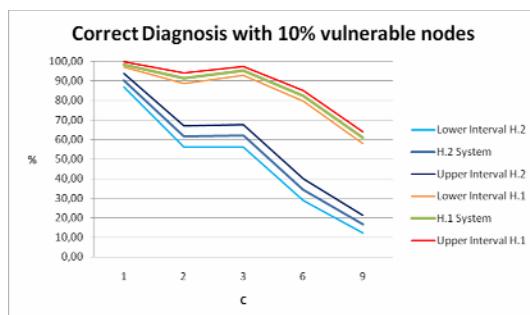


Fig. 3. Results for both systems testing hypothesis H.1 and H.2 with 10% of vulnerable agents in the network in networks of increasing complexity

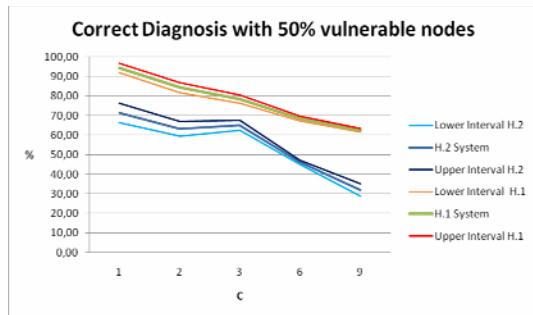


Fig. 4. Results for both systems testing hypothesis H.1 and H.2 with 10% of vulnerable agents in the network in networks of increasing complexity

The results clearly indicate a degradation of the performance with an increase in the complexity of the system. The performance decrease is more accentuated in the system testing hypothesis H.2. In both cases the performance tends to drop more significantly after $C = 3$. The drop in performance is consistent with the results presented in charts 1 and 2. In fact, beyond complexity or connectivity 3 the network denotes a logarithmic behavior in respect to the number of agents affected by the faults, in both cases under test, yielding an higher number of diagnosis performed. When using local information (H.1) the system essentially fails in the determination of the PFO and PFOther once the higher connectivity promotes the propagation through loops and dramatically shortens the average distance between nodes leading to a fault feedback state where is rather difficult to pinpoint both the origin or the end of the fault. On the global (H.2) case the performance drop can be explained due to the learning and progressive reinforcement of rules. Given the random nature of the faults this system tends to perform better when failures are less pervasive. The H.2 systems learns more meaningful rules in this case. The fact that this system is constantly learning also explains why its performance is better when there are more vulnerable nodes in the network since the rule that all the types of agents affect all the type of agents emerges and stabilizes sooner. The type of fault propagation considered in the tests penalizes, as verified, this second system as it becomes much less probable the existence of fault patterns directly mapped to the logic-based diagnostic rules.

7 Conclusions and Outlook

Emerging sustainability challenges are shaping the way in which future shop floors will respond to disturbances and contribute to more rational production patterns. Recent shop floor paradigms have pushed the boundaries of Information Technologies and Artificial Intelligence promoting the integration of components and the seamless reconfiguration of systems. However diagnosis has been left relatively unattended. Current diagnostic approaches are essentially focused in units (either isolated components or entire systems) and some specific parameters. The authors contend that there is added value in considering the interconnectivity of the components as a complementary diagnostic abstraction layer. As the tests expose

systems can behave differently in respect to their network characteristics and catastrophic failures can emerge even for a reduced number of vulnerable nodes. Understanding how these events may develop in the system and being able to explain them in a network of loosely coupled mechatronic agents is a major challenge that has to be tackled if the emerging shop floor control approaches are to become a reality.

The systems presented and tested attempt to capture this network dimension of the diagnostic problem. The tests range from low network complexity to high, simulating distinct relations between components, typically mechanical/physical interactions in the first case and the communication requirements of networks of mechatronic agents in the second. The generic nature of the tests also raises the question that complexity may not be only present in the emerging paradigms as it may also be perceived in current systems (often perceived as more stable given their supporting technologies). Concerning the two hypothesis under test the system validating hypothesis H.1 ranked higher in the tests denoting more stability in the results and less performance degradation. However, assessing the behavior of diagnostic systems in these complex scenarios (from an interaction perspective) is somehow a novelty and it can be argued that the performance of the system testing hypothesis H.2 could be enhanced by using a distinct technology (arguably so could for the case of H.1). One of the secondary goals of the presented tests was, to a certain extent, perceive whether more conventional approaches (logic based reasoning in the present case) would perform, using off-the-shelf technologies, in diagnosing at the proposed abstraction level. In this matter the results suggest the need for tools and approaches closer to the system implemented to test H.1. Excluding the performance advantage, the use of local information allows embedding the diagnostic system at agent level not corrupting the decoupled nature of the underlying control/configuration logic while promoting scalability (fundamental feature of future production systems).

References

- Colombo, A.W.: Industrial Agents: Towards Collaborative Production Automation, Management and Organization. *IEEE Industrial Electronics Society Newsletter* 52, 17–18 (2005)
- Camarinha-Matos, L.M.: Collaborative networked organizations: Status and trends in manufacturing. *Annual Reviews in Control* 33, 199–208 (2009)
- Camarinha-Matos, L.M., Afsarmanesh, H.: Collaborative networks: a new scientific discipline. *Journal of Intelligent Manufacturing* 16, 439–452 (2005)
- Camarinha-Matos, L.M., Afsarmanesh, H.: Elements of a base VE infrastructure. *Computers in Industry* 51, 139–163 (2003)
- Camarinha-Matos, L.M., Afsarmanesh, H., Galeano, N., Molina, A.: Collaborative networked organizations - Concepts and practice in manufacturing enterprises. *Computers & Industrial Engineering* 57, 46–60 (2009)
- Stamatis, K., Colombo, A.W., Jammes, F., Strand, M.: Towards Service-oriented Smart Items in Industrial Environments. *International Newsletter on Micro-Nano Integration - MST News. VDI/VDE-IT*, 11–12 (2007)
- Ueda, K.: A concept for bionic manufacturing systems based on DNA-type information. In: PROLAMAT. IFIP, Tokyo (1992)

8. Babiceanu, R., Chen, F.: Development and applications of holonic manufacturing systems: a survey. *Journal of Intelligent Manufacturing* 17, 111–131 (2006)
9. Koren, Y., Heisel, U., Jovane, F., Moriawaki, T., Pritchow, G., Ulsoy, A.G., Van Brussel, H.: Reconfigurable Manufacturing Systems. *CIRP Annals - Manufacturing Technology* 48, 527–540 (1999)
10. Onori, M.: Evolvable Assembly Systems - A New Paradigm? In: 33rd International Symposium on Robotics Stockholm (2002)
11. Barata, J., Frei, R., Onori, M.: Evolvable Production Systems Context and Implications. In: International Symposium on Industrial Informatics. IEEE, Vigo (2007)
12. WCED: Report of the World Commission on Environment and Development: "Our Common Future" Development and International Economic Co-Operation: Environment. United Nations (1987)
13. Agenda, U.N.: 21. United Nations Conference on Environment and Development. Rio de Janeiro, Brazil (1992)
14. Asif, M., de Brujin, E.J., Fisscher, O., Steenhuis, H.J.: Achieving sustainability three dimensionally. In: 4th IEEE International Conference on Management of Innovation and Technology, ICMIT 2008, pp. 423–428 (2008)
15. Joseph Pine II, B.: Mass Customization: The New Frontier in Business Competition. Harvard Business School Press, Boston (1993)
16. SIRENA: Service Infrastructure for Real-time Embedded Network Applications (2006), <http://www.sirena-itea.org/Sirena/Home.htm>
17. Jammes, F., Smit, H.: Service-oriented architectures for devices - the SIRENA view. In: International Conference on Industrial Informatics, pp. 140–147. IEEE, Perth (2005)
18. Jammes, F., Mensch, A., Smit, H.: Service-Oriented Device Communications Using the Device Profile for Web Services. In: Advanced Information Networking and Applications Workshops, pp. 947–955. IEEE, Los Alamitos (2007)
19. Jammes, F., Mensch, A., Smit, H.: Service-Oriented Device Communications Using the Device Profile for Web Services. In: ACM 3rd International workshop on middleware for pervasive and ad-hoc computing (2005)
20. SODA: Service Oriented Device and Delivery Architecture 2008 (2006 - 2008), <http://www.soda-itea.org/Home/default.html>
21. SOCRADES: Service-Oriented Cross-layer infRAstructure for Distributed smart Embedded devices (2006), <http://www.socrades.eu/Documents/AllDocuments/default.html>
22. InLife: Integrated Ambient Intelligence and Knowledge Based Services for Optimal Life-Cycle Impact of Complex Manufacturing and Assembly Lines (2006), <http://www.uninova.pt/inlife/>
23. EUPASS: Evolvable Ultraprecision Assembly Systems (2006), <http://www.eupass.org/>
24. Isermann, R.: Fault Diagnosis Systems: An Introduction from Fault Detection to Fault Tolerance. Springer, Berlin (2006)
25. Venkatasubramanian, V., Rengaswamy, R., Kavuri, S.N.: A review of process fault detection and diagnosis Part III: Process history based methods. *Computers and Chemical Engineering* 27, 327–346 (2003)
26. Venkatasubramanian, V., Rengaswamy, R., Kavuri, S.N.: A review of process fault detection and diagnosis Part II: Qualitative models and search strategies. *Computers and Chemical Engineering* 27, 313–326 (2003)
27. Amaral, L.A.N., Ottino, J.M.: Complex networks: Augmenting the framework for the study of complex systems. *The European Physical Journal B* 38, 147–162 (2004)

28. Dorogovtsev, S.N., Mendes, J.F.F.: Evolution of networks. *Advances in Physics* 51, 1079–1187 (2002)
29. Watts, D.J.: A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences of the United States of America* 99, 5766 (2002)
30. Ribeiro, L., Barata, J., Ferreira, J.: The Meaningfulness of Consensus and Context in Diagnosing Evolvable Production Systems. In: Camarinha-Matos, L.M., Pereira, P., Ribeiro, L. (eds.) DoCEIS 2010. IFIP AICT, vol. 314, pp. 143–150. Springer, Heidelberg (2010)
31. Ribeiro, L., Barata, J., Ferreira, J.: Emergent Diagnosis for Evolvable Production Systems. In: IEEE International Symposium on Industrial Electronics, Bari, Italy (2010) (accepted)
32. Ribeiro, L., Barata, J., Ferreira, J.: An Agent-Based Interaction-Oriented Shop Floor to Support Emergent Diagnosis. In: IEEE International Conference on Industrial Informatics. IEEE, Osaka (2010) (accepted)
33. Lawrence, R.R.: A tutorial on hidden Markov models and selected applications in speech recognition. *Readings in speech recognition*, pp. 267–296. Morgan Kaufmann Publishers Inc., San Francisco (1990)
34. Ribeiro, L., Barata, J., Alves, B.: Exploring the Network Dimension of Diagnosis in Evolvable Production Systems. In: IEEE Emerging Technologies and Factory Automation (ETFA 2010). IEEE, Bilbao (2010) (accepted)
35. Bonchev, D., Buck, G.A.: Quantitative measures of network complexity. *Complexity in Chemistry, Biology, and Ecology*, 191–235 (2005)

Distributed Accelerometers for Gesture Recognition and Visualization

Pedro Trindade^{*} and Jorge Lobo

Institute of Systems and Robotics

Department of Electrical and Computer Engineering, Faculty of Science and Technology

University of Coimbra – Polo II, 3030-290 Coimbra, Portugal

{pedrotrindade, jlobo}@isr.uc.pt

Abstract. Acceleration information captured from inertial sensors on the hand can provide valuable information of its 3D angular pose. From this we can recognize hand gestures and visualize them. The applications for this technology range from touchless human-machine interfaces to aiding gesture communication for humans not familiar with sign language. The development of silicon chip manufacture allowed these sensors to fit in the top of a nail or implanted in the skin and still wirelessly communicate to a processing unit. Our work demonstrates that it is possible to have gesture recognition from a clutter-free system by wearing very small devices and connect them to a nearby processing unit. This work will focus on the processing of the acceleration information. Methods are shown to estimate hand pose, finger joints' position, and from that recognize gestures. A visualization of the angular pose of the hand is also presented. This visualization can show a single render of the pose of a recognized gesture or it can provide a simple real-time (low-latency) rendering of the hand pose. The processing of the acceleration information uses the gravity acceleration as a vertical reference.

Keywords: Accelerometers; Hand; Gesture Recognition; Gesture Visualization.

1 Introduction

Nowadays the skills of communication are vital to the society. Whether in industrial environments or plain social human activities there is a constant need of good communication skills. In a work environment there may be a need to communicate to a machine in a remote, secure, practical or non-intrusive manner. Such requirements are achieved when it is possible to represent the normal human activity by a virtual representation. In the particular case of human gestures, it is possible to have a virtual representation of the human gesture and be able to communicate in a noisy, possibly at a long distance or with low visibility environment.

When related to human-to-human activities, communication can become complicated when one of the interlocutors does not know the other's language. This is the case when two persons try to communicate and one is hearing impaired and knows sign language, while the other is not and does not know such language. Being possible

^{*} This work is partially supported by the collaborative project HANDLE (grant agreement no 231640) within the Seventh Framework Programme FP7 - Cognitive Systems, Interaction and Robotics.

to intermediate the communication between this two persons allows a great social achievement with an immeasurable value to those who carry that limitation. Science and Industry are constantly evolving and today it is possible to create a system capable of facilitating communications, as mentioned above, and yet be portable, reliable, eventually self-powered and very simple to use. The development of silicon chip manufacture enabled the development of low-cost single chip inertial sensors. By using gravity as a vertical reference these inertial sensors can provide acceleration information from gravity to be used for gestures' angular pose estimation. These sensors can fit in a person's thumbnail or even implanted on the skin. Having them connected to some terminal available to the user, it is possible to have a system available to the user's mobile devices like a smartphone or an iPad.

2 Contribution to Technological Innovation

This work mainly addresses the problem of recognizing simple static sign gestures from the information provided by inertial sensors. By using a triaxial accelerometer in each finger and one on the palm, we can measure the acceleration of gravity in relation to each axis of the sensor. Based on this we are able to estimate the pose of all the hand joints. This enables us to have a clear representation of the hand.

From the representation we developed an algorithm to recognize the gesture against a pre-defined library of gestures. From this algorithm it will be shown that it is possible to have only one reference gesture in the library and still achieve useful results. This avoids a cluster-based approach for recognition and simplifies the process without compromising the results.

Key contributions:

- Study of the information extracted from hand distributed accelerometers.
- Algorithm to estimate the 3D angular pose of each finger, despite of the inobser-vance of rotation in the gravities axis.
- Visual representation in real-time (low latency) and offline of the hand using Py-thon and Blender 3D software package.

3 Related Work

Several types of gesture capture systems are possible. It can be an optical capture system using vision for recognizing the configuration of a hand and it can be a hand sensor- based recognition. Based on the extended analysis of [1], an overview of these systems is given next.

Color markers: Color markers systems for gesture recognition use color patterns to estimate the pose of the hand. This estimation is obtained with inverse kinematics. It has been demonstrated that it is possible to have a low-cost and effective recognition with this approach yet it requires an camera suitably positioned and proper lightning conditions. Our approach only requires minute sensors distributed on the hand.

Bare-hand tracking: These systems typically rely on edge detection and silhouettes and are generally robust to lightning conditions. Reasoning from them involves inference algorithms to search the high-dimensional pose space of the hand. That is computationally expensive and goes far from real-time and becomes unsuitable for Human-Machine interfaces.

Marker-based motion-capture: Marker-based systems involve the use of retro-reflective markers or LED and expensive many-camera setups. Such systems are highly accurate but are expensive. Our proposed system only requires low-cost sensors on the hand.

Data-driven pose estimation: The Data-driven pose estimation makes use of the values from sensors attached to the hand to define the pose of the hand. Such approach allows simple computation to estimate the hand pose. This type of system can easily become very intrusive since it needs the user to somehow wear the sensors. In this work we show that only a few minute non intrusive sensors are enough.

4 Implementation

4.1 Gestures and Portuguese Sign Language

Gestures allow the representation of simple sign language expressions of actions. They can be understood as the static positioning of the fingers in relation to the hand's wrist. It may also include some rotation of the wrist. In Portuguese Sign Language the most elementary words such as the alphabet and the numbers are represented by static gestures. Fig. 1 exemplifies one of these gestures.

4.2 Acceleration Sensors and Motion Classifier

Accelerometers measure linear acceleration, and current sensors integrate three orthogonal sensing axis on a single chip. These can be distributed onto each finger and the palm as shown in Fig. 2.

The sensed acceleration results from gravity and hand motion. In order to use gravity as a vertical reference to infer hand and finger pose, the hand has to be static. The motion classifier looks in the neighborhood of each sample to detect this, assuming sustained steady accelerations of the hand and fingers do not occur.

The modulus of the acceleration vector is used and three thresholds are defined $\{M_1, M_2, M_3\}$. Let W be the depth of neighboring search, for a given sample s_i , its neighbors are defined between $[s_{(i-W)}, \dots, s_{(i+W)}]$ and let L_i be the level of motion of sample s_i , then for each sample s_i

$$L_i = l, M_{l-1} \leq (\max\{s_{(i-W)}, \dots, s_{(i+W)}\} - \min\{s_{(i-W)}, \dots, s_{(i+W)}\}) < M_l, \quad (1)$$

where $l = \{1, 2, 3\}$ and $M_0 = 0$.



Fig. 1. Example of a static gesture from the Portuguese Sign Language

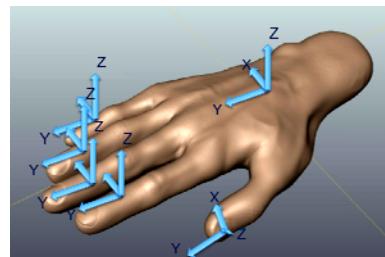


Fig. 2. The different frames of reference for a distributed accelerometer system with a sensor in each finger and the palm

4.3 Gesture Recognition

When there is no acceleration, the gravity vector provides a vertical reference so that the measured acceleration can provide information about the pose of the triaxial accelerometer. However this provides only two inclination angles, but no azimuth since rotations about the vertical are not sensed. To overcome this limitation a method, based on [6] is used, where a method to determine the rotation between two frames of references is proposed by using Horn's [14] closed-form solution for absolute orientation using unit quaternions. Each triad of accelerometers can be seen as an observer of the gravity vertical reference. When the sensor is static the measurements provide a vertical reference vector in the sensor's local frame of reference.

So, by using two set of vectors, one given by the accelerometer on a finger and the other set of vectors given by the accelerometer in the palm it is possible to obtain the full 3D pose of the finger's accelerometer relative to the palm accelerometer.

This results in a feature space of 20-dimensional real vector space (*roll*, *pitch* variable for the thumb and *roll*, *pitch*, *yaw* for the other fingers. A library of gestures is created, so that a nearest neighbor approach can be used to identify an observed gesture. Manhattan distance is used: let p be the observed gesture and q_i the set of gestures in the library, we get a set of distances q_i given by:

$$d_i = \sum_{k=1}^{17} |p_k - q_{i,k}|, \quad (2)$$

and the shortest one indicates the matched gesture.

Fig. 3 shows a general overview for the process of recognizing gestures. The raw sensor output needs some filtering, bias and scale factor compensation. The next step is the motion classifier as explained above. When there is no motion, the process enters the final step of the classifier, where the static gesture is recognized. The detection of sudden motion can also be used to signal the start and end of a gesture for the interface..

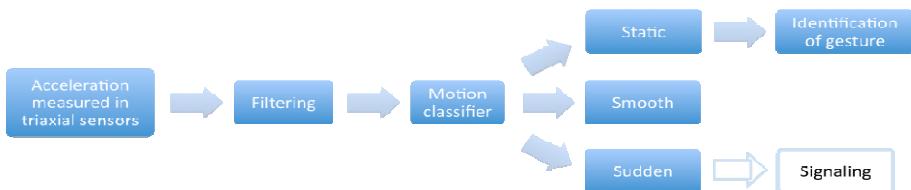


Fig. 3. Overview of the phases for gesture recognition

4.4 Visualization

Relationship between Finger Joints

For the gesture recognition presented above only the fingertip orientation relative to the palm was considered. For rendering a human hand model some assumptions are required. According to [2] the total active range of motion of a typical finger is 260°, in which is the sum of active flexion at the MP (metacarpophalangeal) (85°), PIP (proximal interphalangeal) (110°), and DIP (distal interphalangeal) (65°) joints. The

range of active extension at the MP joint varies between people but can reach 30- 40° [3]. Although there is a considerable axial rotation in addition to flexion/extension and abduction/adduction movements, this is constrained and so it is not considered a true third degree of freedom. In total, the human hand, including the wrist has 21 degrees of freedom of movement [2][4]. A common reduction of the number of degrees of freedom is to consider that for the index, middle, ring, and little fingers, in order to bend the DIP joints the correspondent PIP joints must also bend, as seen in Fig. 4. According to [4] the relationship can be approximately presented as:

$$\theta_{DIP} = \frac{2}{3} \times \theta_{PIP} \quad (3)$$

Software

3D graphics is the grandchild of Euclid's Elements, a geometric construction of the Universe as a mesh of points connected by measurable lines [7]. OpenGL is a natural choice for drawing those primitives. But to allow more control and fast implementation, a higher level software layer is needed. Blender is a free open source 3D content creation suite that allows several different approaches for 3D representation. Moreover, it allows scripting control of the software. Also, choosing Blender allowed the use of this visualization tool for datasets of the European project HANDLE¹ [8].

Blender allows Python scripting language to control all the environment of the software. This means it is possible to dynamically use one's own equations and algorithms to set up the environment. This resulted in reading the output of the processed information from other external sources (such as Matlab or any other program) and have this information resulting in a user-controlled rendering. This workflow can be seen in Fig. 5.

As indicated in Fig. 5 we use Matlab for the offline processing for gesture recognition and Python for online processing the accelerometer data and in both we update a file that is continuously monitored by Blender to update the rendered model.

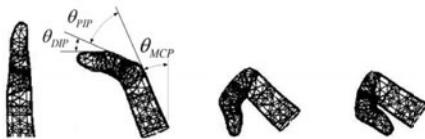


Fig. 4. Constraints between PIP and DIP joints

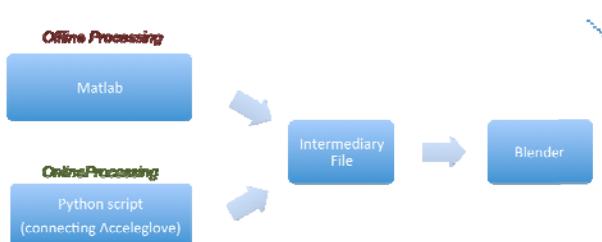


Fig. 5. Workflow for the visualization of the hand angular pose

¹ HANDLE is a collaborative project funded by the European Commission within the Seventh Framework Programme FP7, as part of theme 2: Cognitive Systems, Interaction, Robotics, under grant agreement 231640.

5 Results

5.1 Experimental Setup

To implement the methods in section 4, Matlab and Python were used. Matlab for the processing and Python to interface with Blender.

For the distributed accelerometer system the Anthrotronix Acceleglove [9] was used. In the Acceleglove the accelerometer used is the MMA7260QT from Freescale Semi-conductor [5]. This device is a low-cost capacitive micromachined accelerometer and can measure acceleration up to 6G.

5.2 Recognition

From Fig. 6 it is possible to see the motion level classifier working. According to the deviation in the modulus of each sensor a classification of the level of movement is made, meaning Level-1 to be fairly static, while Level-2 refers to a smooth movement and Level-3 refers to a sudden movement that eventually could be used to signal the start and stop of a stream of gestures.

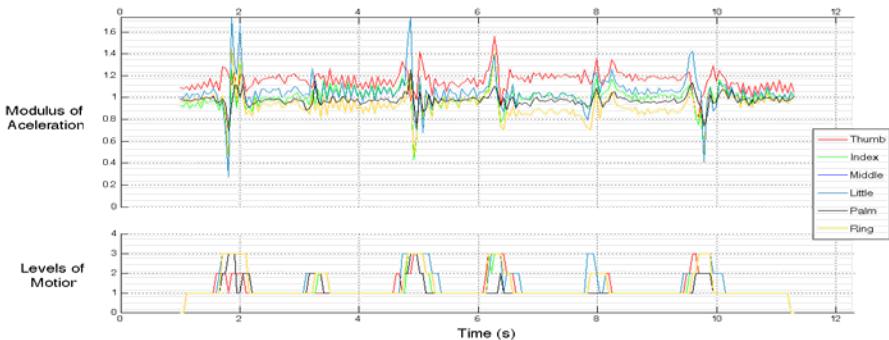


Fig. 6. Example of the motion classification after a data acquisition

Each contiguous subset of Level-1 samples vectors are converted into a single vector, called frame. All the frames are then used to find the quaternion of rotation between each finger and the palm. After calculating the quaternion of rotation a reprojection error is measured for each of the frames. This reprojection error is shown in Fig. 7.

The values found on Fig. 7 allow saying reprojection error is very small, with all

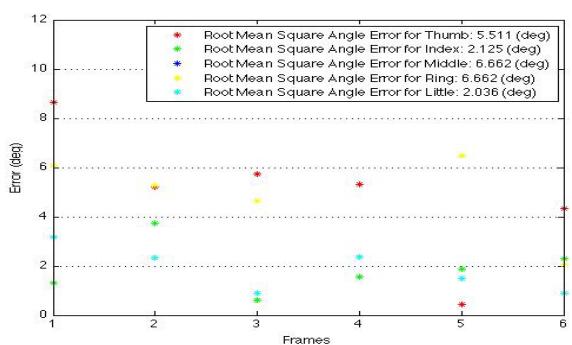


Fig. 7. Rotation reprojection error after the calculation of the quaternions of rotation

the values standing before 7 degrees. This correspond to 3 distinct positions (frames) for which the user had to maintain the same gesture, and a few degrees will not significantly alter the gesture so this result indicates that the estimated relative angular pose values are suitable for gesture recognition.

The gesture that it is being analyzed is compared against a library. In this library each entry relates all the information of a single gesture. That includes the name of gesture, the roll, pitch and yaw angles, the quaternion values and weight that define how much impact should roll, pitch or yaw values have in the recognition. Comparing a gesture against the library produced the results like the ones shown in Table 1.

Table 1. Manhattan distances from the current gesture to the ones in the library

Gesture	G	H	K	L	O	Q	R	S	T	U	V	W
Distance	6.0	3.1	4.2	3.9	4.2	5.2	5.3	5.6	4.3	4.2	7.0	4.4
Gesture	X	Y	Z	1	2	3	5	4	6	7	8	9
Distance	7.4	4.0	5.0	6.0	3.9	6.2	5.5	6.3	4.3	6.4	5.1	4.2

In Table 1 it is possible to see a list of all the gestures included in the library and the distance, in a Manhattan geometry of the current gesture to the ones in the library.

The gesture performed that resulted in the comparison shown in Table 1, was indeed the “H” gesture. The visual perspective of the gesture is shown in Fig. 1.

5.3 Visualization

The visualization was structured to allow the representation of the hand pose processed in the recognition process and also directly from a real-time connection to the Aceleglove [9]. The visualization of the results that came from Matlab was possible to represent because Matlab outputted the Roll Pitch and Yaw values for each sensor it processed. Having blender prepared to read those values and running the Python routines defined in blender a correct pose from blender was possible to represent. Fig. 8 shows a render of the pose. Again, the relatively small error in calculating the hand pose allowed the visual representation as seen in this figure.

When this visualization was performed in real time, an external (to Blender) script was running. This script would provide Blender the pose information for each sensor. At a framerate of about 29fps, Blender updates the pose of every joint in the hand and renders it. This performance was achieved in a Macbook Pro 2.5GHz from 2009.

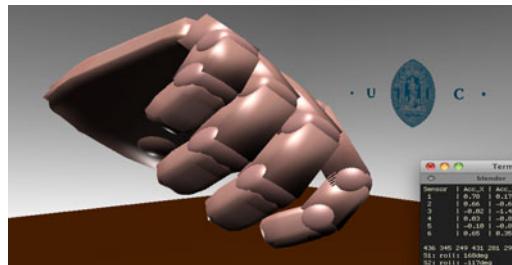


Fig. 8. A render from Blender showing the encountered angular pose of the hand

6 Conclusions

The recognition of gestures is a vast area of research. In this work a novel approach using distributed accelerometers in the hand to recognize the gestures solely by the measurement of acceleration was presented.

By using an intelligent distribution of accelerometers on the hand it was possible to create an algorithm to recognize a hand gesture. Our approach relies on the vertical reference provided by gravity to infer angular pose, however in this way rotations about vertical axis are not observable. To overcome this limitation, a novel approach, based on the work of [6] on relative pose calibration between inertial sensors and cameras, allowed to find the exact 3D angular pose of the fingers in relation to the palm. The recognition is based on observation of multiple frames of static gestures. As future work this recognition should be extended to allow dynamic gesture to be recognized as well.

We also proposed a 3D visualization tool for the hand pose. By creating an intelligent structure this tool was already capable of fulfilling a broader project, like the HANDLE project [8]. The approach presented in this work can be improved in many ways, by addressing dynamic gestures, applying a probabilistic approach for learning the gestures, or even using more accelerometers on the intermediate joints.

References

1. Wang, R., Popovic, J.: Real-Time Hand-Tracking with a Color Glove. ACM Transaction on Graphics (2009)
2. Jones, L., Lederman, S.: Human Hand Function. Oxford University Press, Oxford (2006)
3. Kapandji, I.A.: The Physiology of the joints: Upper limb, 2nd edn. E & S Living- stone, Edinburgh (1970)
4. Li, K., Chen, I.-M., Yeo, S.H.: Design and Validation of a Multi-finger Sensing Device Based on Optical Linear Encoder. In: IEEE International Conference on Robotics and Automation. Anchorage, Alaska, USA (2010)
5. Freescale Semiconductor MMA720QT webpage, http://www.freescale.com/webapp/sps/site/prod_summary.jsp?code=MMA7260QT
6. Lobo, J., Dias, J.: Relative Pose Calibration Between Visual and Inertial Sensors. Int. J. Rob. Res. 26, 561–575 (2007)
7. Oliver, J.: Buffering Bergson: Matter and Memory in 3D Games (2008)
8. HANDLE European Project website, <http://www.handle-project.eu>
9. AnthroTronix Acceleglove, <http://www.acceleglove.com/>
10. Barbour, N.M.: Inertial Navigation Sensors, Sensors & Electronics Technology Panel. Draper Report no. P-4151 (2003)
11. Baudel, T., Beaudoin-Lafon, M.: Charade: remote control of objects using free-hand gestures. Communications of the ACM 36(7), 28–35 (1993)
12. Fraden, J.: Handbook of Modern Sensors: Physics, designs, and applications, 3rd edn. Springer, Heidelberg (2003)
13. Ferreira, A., et al.: Gestuário - Língua Gestual Portuguesa. Secretariado Nacional para a Reabilitação e Integração das Pessoas com Deficiência (1997)
14. Horn, B.K.P.: Closed-Form Solution of Absolute Orientation Using Unit Quaternions. Journal of the Optical Society of America 4(4), 462–629 (1987)

15. Khan, S., Gupta, G., Bailey, D., Demidenko, S., Messom, C.: Sign Language Analysis and Recognition: A Preliminary Investigation. IVC New Zealand (2009)
16. Takayuki, H., Ozake, S., Shinoda, H.: Three-Dimensional Shape Capture Sheet Using Dis-tributed Triaxial Accelerometers. In: 4th Int. Conf. on Networked Sens. Syst. (2007)
17. Taylor, C.L., Schwarz, R.J.: The anatomy and mechanics of the human hand. Selected Articles From Artificial Limbs, 49–62 (1970)
18. Weinberg, M.S., Kourepinis, A.: Error Sources in In-Plane Silicon Tuning-Fork MEMS Gyroscopes. Journal of Microelectromechanical Systems 15(3) (2006)

Towards Statecharts to Input-Output Place Transition Nets Transformations

Rui Pais^{1,2,3}, Luís Gomes^{1,2}, and João Paulo Barros^{2,3}

¹ Universidade Nova de Lisboa, Faculty of Sciences and Technology, Portugal
`{ruipais@uninova.pt}`

² UNINOVA, Center of Technologies and Systems, Portugal
`{lugo@fct.unl.pt}`

³ Instituto Politécnico de Beja, Escola de Superior Tecnologia e Gestão, Portugal
`{jpb@uninova.pt}`

Abstract. This paper proposes a set of procedures addressing a Model Driven Architecture approach to translate of SysML statechart models into a class of non-autonomous Petri nets. The main goal of this set of procedures is to benefit from the model-based attitude allowing the integration of development flows based on statecharts with the ones based on Petri nets.

Several methodologies exist to transform statechart models into specific classes of Petri net models, which depend on the proposed goals to achieve. The target formalism for the translation is the class of Input-Output Place Transition Nets, which extends the well-known low-level Petri net class of place transition nets with input and output signals and events dependencies. With this Petri net class we aim to contribute with tools to be integrated on a framework for the project of embedded systems using co-design techniques.

Keywords: Statecharts, Petri Nets, SysML, PNML, MDA, ATL.

1 Introduction

Systems engineering problems are becoming increasingly complex. Model-Driven Software Development (MDSD) is recognized as an auspicious approach to deal with software complexity. As mentioned by Gregor Engels [1], the main goal of Model-Driven Architecture approach is to obtain the automatically generated code from behavioral models. In most cases, behavioral models are only used on early stages of a software development project to document user requirements, many times created from uses cases models. Later, they are used on the first implementation steps as a support to identify user and systems requirements. However, during requirements and code changes, behavioral models quick stays unconscious once it is not considered valuable its maintenance.

Object Management Group's (OMG) Model Driven Architecture (MDA) provides the basic terminology for MDSD. Software Development under MDA is a new software development paradigm. The vision of MDA aims to promote modeling to a central role in the development and management of application systems, permitting fully-specified platform-independent models (including behavior), decoupling the way

that application are defined from the technology they run on. This should improve that investments made in building systems can be preserved even when the underlying technology platforms change.

Considering currently available MDA transformation tools, it was decided that development should use the Eclipse Model-to-Model Transformation (M2M) project in conjunction with ATLAS Transformation language (ATL), which is a Query/View/Transformation like transformation language, and provides an open source development under Eclipse Generative Modeling Technologies (GMT) project.

The main goal that we intend to archive in the near future, using MDA, is the translation of behavioral Metamodels to an intermediate Metamodel based on Petri Nets (PN). This transformation will permit analyze of equivalent Petri Net properties, both static concepts (conflicts, priorities, etc.) and their dynamic interpretation (liveness, enabledness, fireability). From the Petri Nets Metamodel, it will be also possible to get support for automatic code generation for its simulation and execution.

The behavioral model in analysis is the Unified Modeling Language (UML) Statecharts, which is a state machine variant having its origins in the well-known formalism introduced for the first time by Harel in [2], with its static semantics being described by Metamodels and a constraint language. UML Statecharts are well-known design methodology to capture the dynamic behavior of reactive systems, that helps specify, visualize, and document models of software systems.

The Object Management Group has recently developed the Systems Modeling Language (SysML), which supports the specification, analysis, design, verification and validation of a broad range of complex systems. It permits the modeling of processes embracing software engineering, mechanics, electrics and electronics areas. SysML extends UML (Unified Modeling Language) and is adapted to model systems which are not entirely software based. SysML included UML state diagrams that are essentially a Harel statechart.

The remainder of this paper is structured as follows: Section 2 presents motivation and innovations; section 3 mentions related work; section 4 introduces the Petri Net class used "Input-Output Place Transition class" (IOPT); in section 5 is presented translations rules to convert Statecharts elements into correspondent IOPT items; section 6 illustrates the transformation from Statecharts to IOPT Nets using as example a controller for a simple Railway System. Section 7 winds up with conclusions and future work.

2 Contribution to Sustainability

Peter Sandborn defines sustainability as "keeping an existing system operational and maintaining the ability to manufacture and field versions of the system that satisfy the original requirements" [10].

Sustainability is a general term usually referring to environmental, business, and technological sustainability. Since all these visions of a system are interconnected, a positive change in a single activity can contribute to increase the global system sustainability. For example, the development of better system functionality can reduce human effort, improve product or business incomings, etc..

The present work global mission relates to the reduction of system development and testing time through automated code generation based on the analysis and

translation of SysML models. This development will significantly contribute to increase systems sustainability, as it allows a reduction of the development, test and maintenance time, as well as the software technology obsolescence. It also permits the increase of corporate productivity. More specifically, UML and SysML can be used to model the architecture and behavior of a complex system, allowing different points of view. Transformation of behavior models (activity diagrams, statecharts, use case diagrams, interaction diagrams, and others) into specific classes of Petri net models (autonomous models, non-autonomous models, stochastic, timed, etc.) is a well-known problem. The strategies used on models transformation significantly differ in terms of the used source model, target model, restrictions on the models, programming language, used algorithms, and goals to be achieved.

In summary, the following are some of the reasons to perform formalism transformations:

- Code generation - Generation code permits the creation of tools to create, edit, check, optimize, transform and generate simulator for its execution or visualization; Code can be generated for different target platforms and languages (C, SystemC, VHDL, etc.);
- Model manipulation - possibility to apply well-known transformations on Petri Net models; transformations to reduce model complexity; possibility of model reorganization, division, and other manipulations;
- Properties analysis - Analysis and verification of model properties and constraints;
- Precise formalization for the behavior of UML or SysML diagrams.

3 Related Work

A transformation between State Machine diagrams and Time Petri Nets is presented on [8] with the objective of analysis and verification of embedded real-time systems with energy constraints. At first sight this approach has some similarities but with a totally different goal and results. A Multi-Paradigm approach to the modeling of complex systems is presented on [7] where a tool AToM3 is presented (similar to a MDA transformation tool) and an example is presented of transforming Statechart models (without hierarchy) into behaviorally equivalent Petri-Nets. Its focus is in AToM3 tool and not in the transformation example. In [9], translation of hierarchies and other constructs in Statecharts to a specific class of Petri nets is presented, however no formalization neither tool support are referred.

On our vision, the ultimate goal of the transformations between behavioral models and Petri Nets is to generate optimized code to model execution. To achieve this goal, many steps will be needed, where all the others transformations can generate useful information and equivalents formalism to be considered. Petri Nets have in this process an important role, once there are many researches on Petri Nets models simplification, transformation, code generation and Petri Net Classes.

The transformation of Statecharts to Input-Output Place Transition Nets will be integrated in a framework in development that will be able to convert different behavior models to a common specification based on Petri Net Classes, and from

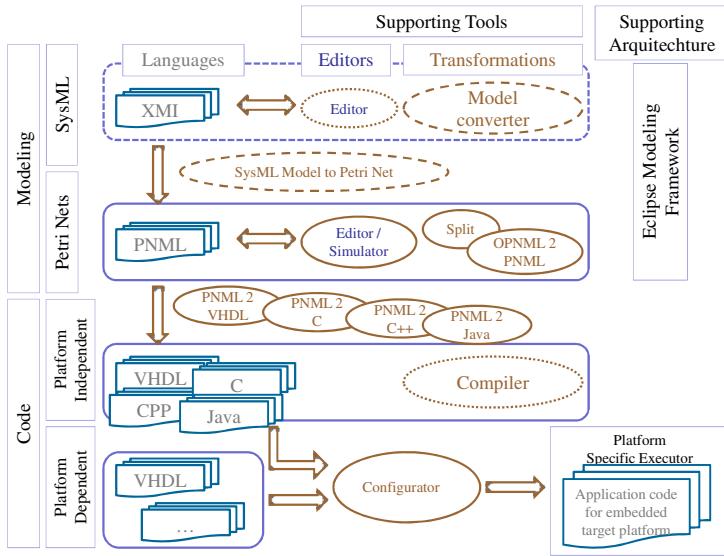


Fig. 1. Framework in development

there, platform independent and dependent code will be generated to create a model executor able to run on specific platforms. Figure 1 illustrates the framework where this contribution will be integrated, on “SysML Model to Petri Net” process. Several others contributions will be done and integrated in this process, creating a very valuable tool to be used on the modeling development process.

4 The Input Output Place Transition Net Class

The Input-Output Place Transition class (IOPT) extends place-transition nets with non-autonomous constructs. It is a class of non-autonomous Petri nets in the tradition of interpreted and synchronized Petri nets of Moalla et al. [3], Manuel Silva [5], and David and Alla [4], named Input-Output Place Transition nets (IOPT) and proposed in [6].

IOPT nets get their name from the possibility to explicitly model external input and output events and signals. The events impose an environment dependency on the Petri net model. More specifically, a transition can be fired only if it is *enabled* and *ready*. As usually, a transition is *enabled* depending on the marking of its input places. Yet, it is *ready* depending on an additional guard defined as a function of external input signals, as well as on input events. Additionally, IOPT nets have a stepwise maximal firing semantics: in each step, all transitions that are *enabled* and *ready* are fired. The IOPT syntax and semantics, as well as the respective rationale, were already formally presented in [6]. Compared to place-transition nets, IOPT nets have the following additional characteristics: (1) Input and output events and signals with an edge level for input events; (2) Two types for input and output signal values; (3) Test arcs and arc weights in normal and test arcs; (4) Priorities and input signal guards in transitions; (5) Each transition can have a set of associated input events and a set of associated output events; (6) Each place can have a set of output signal conditional

assignments and a bound attribute; (7) An explicit specification for sets of conflicting transitions (conflict sets) and sets of synchronous transitions (synchronous sets).

The present work benefits from the results of the FORDESIGN project [4] where a set of tools were developed allowing the use of IOPT nets for code generation from models, namely the following: (1) a graphical editor (including animation capabilities); (2) a tool for the textual specification of model compositions (OPNML2PNML); (3) a tool for the decomposition of an IOPT model into a set of concurrent submodels; (4) translators to C and VHDL, allowing automatic code generation; (5) a configurator (for generating final code considering a specific hardware-software platform). Those tools rely on the PNML interchange format and on Petri net Type definition defined by a Relax NG grammar. Currently, we foresee the use of an Ecore metamodel for the IOPT class, presented in [10], thus offering improved robustness and maintainability, especially for a new generation of code generators.

5 Translating Statecharts to Input Output Place Transition Nets

This section presents translations rules to convert statecharts elements into the correspondent Input Output Place Transition Nets items. The proposed translation procedures are based on the analysis of specific situations, through the analysis of specific model characteristics.

On the current development stage, only a selected set of statecharts elements are considered: states, transitions, events, guards, action, as well as composite states, as orthogonal states (and-sets) and mutually exclusive states (xor-clusters). Other relevant items for specific modeling situations are not discussed in this paper, namely communication mechanisms, hierarchical structuring, preemption, history mechanisms, and others. Yet, the referred subset of selected characteristics is adequate to model a large number of systems. In the next section, the validation of their application to a specific application is presented. In the following subsections translation techniques for the referred elements are presented.

5.1 Mapping States

Statecharts are basically constructed from state diagrams. A state represents a situation or context in a given time instant. States can be classified into simple states (normal state, initial state, or final state), and composite states, which can be of two kinds: the *and-set* and the *xor-cluster*. The xor-cluster is a state machine, in the sense that it is composed by a set of states and transitions, where at most one state can be active at a specific point in time. The and-set is the parallel composition of a set of xor-clusters, where all the xor-clusters in the and-set are active or inactive at a specific instant. In this sense, a simple state is a state that does not have any substates and corresponds to the final level of refinement. Composite states are substates that were refined through the decomposition or abstraction process. The and-set is a composite state with one or more regions. A region is simply a container for substates (a state diagram). Concurrent states correspond to states belonging to different regions that are executed concurrently.

A simple statechart state is translated to a simple IOPT place with the same name, or with a new automatic unique name if the initial statechart state does not have one, as illustrated in Figure 2.

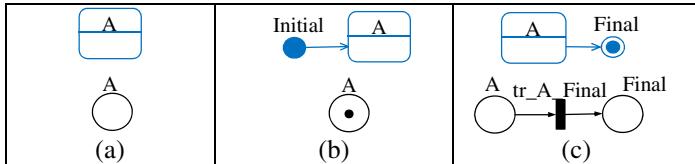


Fig. 2. (a) Statechart normal state translation; (b) Statechart initial state translation; (c) Statechart final state translation

Initial states are active the first time the model is run. The initial state is pointing to another state through a transition. This block (initial state + transition + state) will be translated into a Petri Net place marked with a token corresponding to the destination state, as illustrated in Figure 2b.

The statechart final state corresponds to a state from which its execution can no longer evolve. After the execution of the state previous to the final state, the system should evolve to a state without further evolutions. Consequently, the final state will be translated to a PN final place, a place without output transitions, as illustrated in Figure 2c.

5.2 Mapping Output Actions Associated to States

States can have activities associated with it and internal transitions. Activities can be from one of three types: *entry*, *do*, and *exit*. The *entry* activity is performed when the state becomes active; the *do* activity is performed as long as the state is active; the *exit* activity is performed when leaving the state (which means that the state becomes not active); finally, the internal transitions are events that may occur without causing state changes.

At the current stage, only do activities associated to a state are considered. They will be translated as output signals associated to a place, as illustrated in Figure 3. Translation of entry and exit activities will be part of future work.

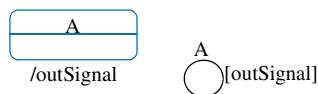


Fig. 3. Statechart activity translation

5.3 Mapping Transitions

A transition is a relationship between two states; upon the firing of a transition the first state will become non active while the second state will become active. Transitions can be classified into *simple*, *join*, or *fork* transition.

A simple transition connects a source state to a target state, and is translated by a Petri Net transition connected by the places that correspond to source and target states, as illustrated in Figure 4.

In the general case, the source and target state of a transition may be at a different level in the state hierarchy.

5.4 Mapping Inputs and Outputs Dependencies Associated with Transitions

Transitions are supposed to represent actions, which are atomic (not interruptible). A transition can have an associated triple (a set of input events, a Boolean guard and a set of output actions) “**Event[Condition]/Action**” where all parts of this triple are optional.

A transition is defined as enabled, if it can be fired. A transition can be fired if its source state is active in the current configuration, its event is present, and its guard is satisfied. When it is fired, the source state is left, the transition actions are executed, and the target state is entered. The technique is illustrated in Figure 5.

5.5 Mapping Orthogonal States

Orthogonal states are mapped as states, which mean that each and-set, all composing xor-clusters and associated states are translated into places. Fork vertices serve to split an incoming transition into two or more transitions terminating on orthogonal target vertices (i.e., vertices in different regions of a composite and-set). The segments outgoing from a fork vertex must not have guards or triggers. The technique is illustrated in Figure 6. Join vertices serve to merge several transitions coming from source vertices in different orthogonal regions. The transitions entering a join vertex cannot have guards or triggers. The technique is illustrated in Figure 7.



Fig. 4. Statechart Simple Transition translation

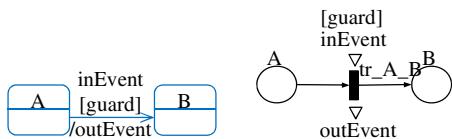


Fig. 5. Statechart Transition Events and Guard translation

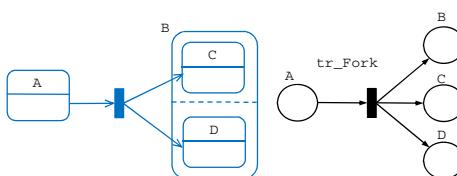


Fig. 6. Statechart Fork Transition translation

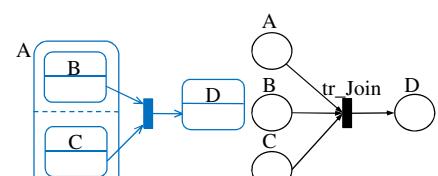


Fig. 7. Statechart Join Transition translation

5.6 Mapping Mutually Exclusive States

A xor-cluster contains mutually exclusive states, which are literally like embedding a statechart inside a state. Two simple situations are considered: when the xor-cluster has no final state, and when the xor-cluster has a final state. In both situations both the xor-cluster and associated states are translated into places, as before. Starting with the former situation, when no final state is present, the proposed translation techniques, already proposed for states and transitions, are still valid, and will be complemented by a explicit modeling

of the outgoing transition which will be replicated for all internal states. The proposed technique is illustrated in Figure 8, where the transition from state B to state E is translated by two transitions (as many as the number of states of the xor-cluster B), referred as tr_Join and tr_JoinBC_E. In order to avoid conflicts, priorities are associated with transitions, as illustrated in the IOPT net in the bottom of Figure 8.

6 A Case Study

To illustrate the transformation from statecharts to IOPT Nets, the example of the controller for a simple Railway System is used. This example includes a scenario where there is a crossing of a railway line, allowing the movement of trains in both directions (one direction at a time), with a highway motor vehicles and a gate, which allows crossing isolation when the train is passing. The system is shown in Figure 9. The

statechart model is presented in Figure 10 and the respective IOPT net model in Figure 11. The initial state of the system admits that the railway line is clear and that the gate arm is open.

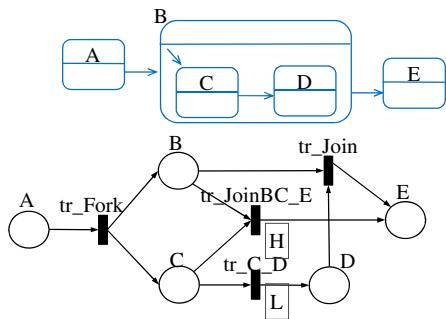


Fig. 8. Statechart and the respective IOPT net with Mutually Exclusive States

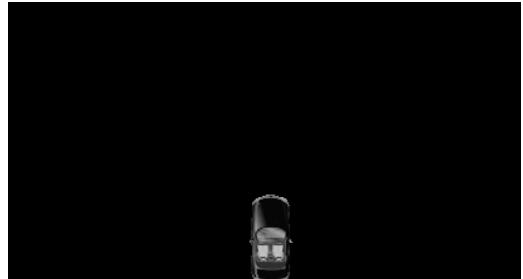


Fig. 9. Railway System

7 Conclusions and Future Work

We have presented a proposal for the translation from statecharts to a class of non-autonomous Petri nets, the IOPT nets. This class of nets is able to model external input and output signals and events, thus allowing the alternative use of Petri nets. Besides the Petri nets known advantages, the IOPT nets allow the use of modular

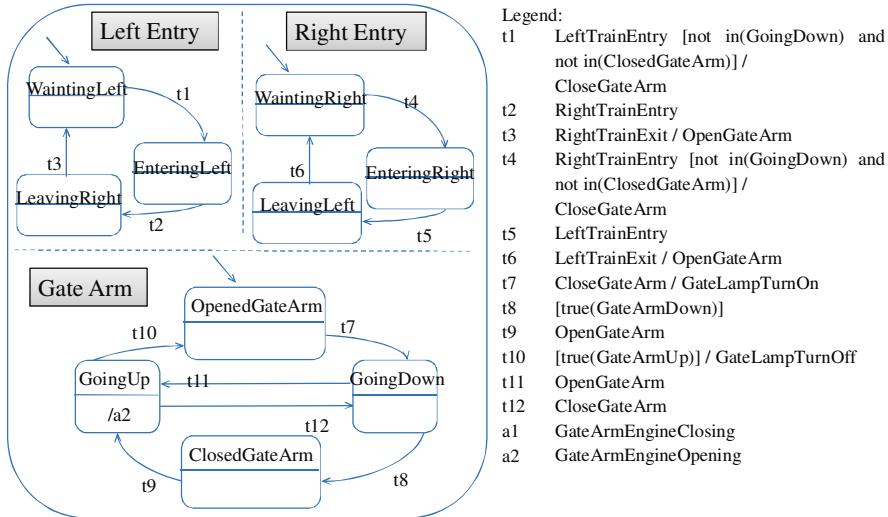


Fig. 10. Railway System Statechart

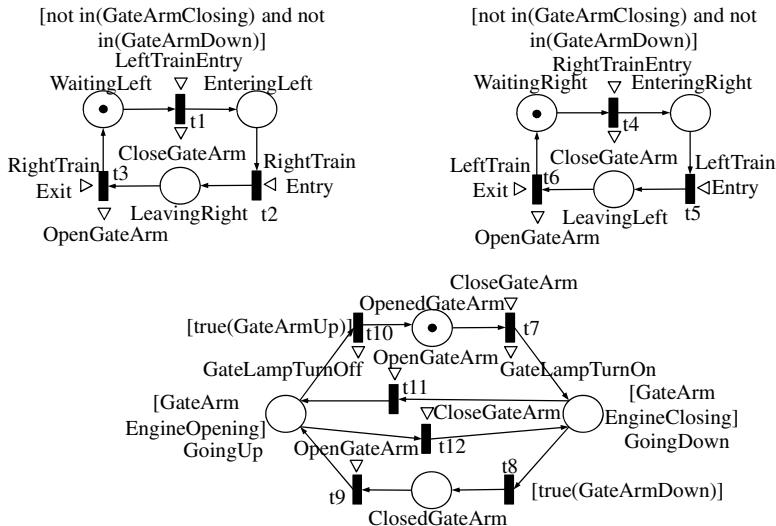


Fig. 11. Petri Nets corresponding to each of the three subcharts in Figure 10

constructs and can be used as an interchange format among a set of tools. The IOPT Ecore metamodel introduces IOPT nets in the MDA infrastructure, allowing MOF simulation. Hence, it becomes possible to interchange data between IOPT tools and other MDA tools by reusing the associated XMI representation, and to support automatic code generators and the exploration of the automatic proofs of behavioral equivalence between two models, using successive transformations (bidirectional transformations) and its simulation.

References

1. Engels, G.: Keynote: Automatic generation of behavioral code - too ambitious or even unwanted?, Behaviour Modelling in Model Driven Architecture. In: Proceedings of First European Workshop on Behaviour Modelling in Model Driven Architecture (BM-MDA), Enschede, The Netherlands, June 23 (2009)
2. Harel, D.: Statecharts: A visual formalism for complex systems. *Science of Computer Programming* 8(3), 231–274 (1987)
3. Moalla, M., Pulou, J., Sifakis, J.: Synchronized Petri nets: A model for the description of non-autonomous systems. In: Winkowski, J. (ed.) MFCS 1978. LNCS, vol. 64, Springer, Heidelberg (1978)
4. David, R., Alla, H.: Petri Nets & Grafset; Tools for Modelling Discrete Event Systems. Prentice Hall International, UK (1992)
5. Silva, M.: Las Redes de Petri: en la Automática y la Informática. Edit. AC, Madrid (1985)
6. Gomes, L., Barros, J., Costa, A., Nunes, R.: The Input-Output Place-Transition Petri Net Class and Associated Tools. In: Proceedings of the 5th IEEE International Conference on Industrial Informatics, INDIN 2007 (2007)
7. de Lara, J., Vangheluwe, H.: Computer Aided Multi-Paradigm Modelling to Process Petri-Nets and Statecharts. In: Corradini, A., Ehrig, H., Kreowski, H.-J., Rozenberg, G. (eds.) ICGT 2002. LNCS, vol. 2505, pp. 239–253. Springer, Heidelberg (2002)
8. Carneiro, E., Maciel, P., et al.: Mapping SysML State Machine Diagram to Time Petri Net for Analysis and Verification of Embedded Real-Time Systems with Energy Constraints. In: Proceedings of the 2008 International Conference on Advances in Electronics and Micro-Electronics, pp. 1–6 (2008) ISBN:978-0-7695-3370-4
9. Gomes, L.: As Redes de Petri Reactivas e Hierárquicas - integração de formalismos no projecto de sistemas reactivos de tempo-real (in Portuguese); PhD Thesis, UNL-FCT (1997), <http://hdl.handle.net/10362/2560>
10. Moutinho, F., Gomes, L., Ramalho, F., Figueiredo, J., Barros, J.P., Barbosa, P., Pais, R., Costa, A.: Ecore Representation for Extending PNML for Input-Output Place-Transition Nets. In: 36th Annual Conference of the IEEE Industrial Electronics Society, IECON 2010, Phoenix, AZ, USA, November 7-10 (2010)
11. Sandborn, P., Myers, J.: Designing Engineering Systems for Sustainability, CALCE, Department of Mechanical Engineering. University of Maryland, <http://www.enme.umd.edu/ESCML/Papers/SustainmentChapter.pdf>

Petri Net Based Specification and Verification of Globally-Asynchronous-Locally-Synchronous System

Filipe Moutinho^{1,2}, Luís Gomes^{1,2}, Paulo Barbosa³, João Paulo Barros^{2,4}, Franklin Ramalho³, Jorge Figueiredo³, Anikó Costa^{1,2}, and André Monteiro³

¹ Universidade Nova de Lisboa, Faculdade de Ciências e Tecnologia, Portugal

² UNINOVA, Portugal

{fcm, lugo, jpb, akc}@uninova.pt

³ Universidade Federal de Campina Grande, Brazil

{paulo, franklin, abrantes, andre}@dsc.ufcg.edu.br

⁴ Instituto Politécnico de Beja, Escola Superior de Tecnologia e Gestão, Portugal

Abstract. This paper shows a methodology for Globally-Asynchronous-Locally-Synchronous (GALS) systems specification and verification. The distributed system is specified by non-autonomous Petri net modules, obtained after the partition of a (global) Petri net model. These modules are represented using IOPT (Input-Output Place-Transition) Petri net models, communicating through dedicated communication channels forming the GALS system under analysis. This set of modules is then automatically translated into Maude code through a MDA approach. As the modules of GALS systems run concurrently, the Maude semantics for concurrent objects is used along with message representation. Finally, as a particular case, the system state space is generated from the Maude specification of the GALS system, allowing property verification.

Keywords: GALS, Embedded Systems, Petri Nets, Maude, Verification.

1 Introduction

Embedded systems are increasingly present in people's lives, for instance in people's pockets, homes, cars and in industrial machinery. Many embedded systems are synchronous systems implemented in a single device, but there are embedded systems that can not be implemented using the synchronous paradigm, either due to the need of having multiple devices in different physical locations, or due to the simple fact that a single device is not enough to implement the system, or even when it is necessary to use devices containing multiple clock domains.

These systems are Globally-Asynchronous-Locally-Synchronous (GALS). They support features like asynchronous messaging and multiple concurrent synchronous modules with different clock domains. But these features make GALS systems development not a simple task, and with greater challenges (for example the verification of GALS systems components interaction) compared to the development of synchronous systems. This was the environment where was found the research

question of this work, which is *How to specify, simulate, verify and implement a GALS system described through a set of IOPT Petri net sub-models supported by automatic code generation?*

The methodology proposed here to develop GALS systems for embedded systems applications was initially formulated within the FORDESIGN project [1], which had the objective to center the development effort in the system modeling, relying in a model-base development attitude and taking advantage of automatic code generation tools. However, the FORDESIGN project did not fully consider the development of GALS systems. The chosen modeling formalism was the Input-Output Place-Transition (IOPT) Petri net, a class of non-autonomous Petri nets defined in [2] that extends the well-known Place-Transition (P/T) Petri net class with inputs and outputs signals and events (among other characteristics).

The Net Splitting Operation proposed in [3], allowing model partitioning into several components, is here used to split centralized models of GALS systems into GALS components. These components will be interconnected through lossless communication channels with undetermined propagation time for all the messages.

To allow property verification of the GALS systems, modeled by IOPT net models, are performed a set of transformations from the IOPT net representation to Maude specifications [4]. Those transformations rely on a Model-Driven Architecture (MDA) [5] approach, using IOPT nets and Maude metamodels. The resulting Maude specifications support the verification of several properties.

The reference development methodology fully integrate design automation tools, namely the PNML2C [6] and PNML2VHDL [7] tools, which automatically generate, respectively, C or VHDL code from IOPT net models represented in the PNML format [8].

Section 2 briefly presents the technological contribution of the paper to sustainability. Section 3 presents the proposed methodology with the help of a running example that will be used throughout the paper for an easier understanding of the development flow. Section 4 describes the executable semantics of GALS systems. Section 5 briefly explains the GALS System representation in Maude language, and the running example verification. Section 6 gives an overview of some related work. And finally section 7 presents some final remarks.

2 Contribution to Sustainability

This work aims to contribute to the development of GALS systems in a more automated way relying in the Model-Driven Architecture (MDA) approach. This allows the development of complex systems in less time, while being more reliable and less vulnerable to development bugs, due to the fact that the only development errors that are introduced in the system are the modeling errors; there is no manual code generation, either simulation, verification, or implementation codes. Systems with bugs can cause unwanted effects, as in most cases they need to be replaced, wasting energy and resources.

3 Proposed Methodology

The development flow proposed to GALS systems behavior verification comprises the following steps (described in Fig. 1):

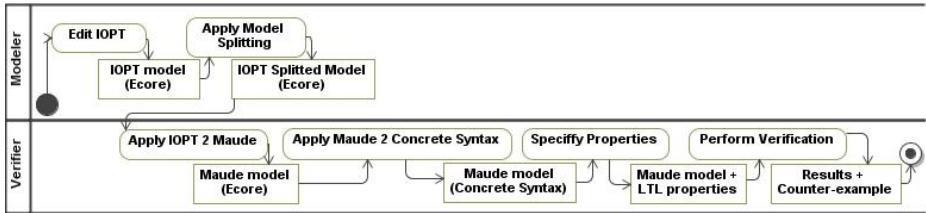


Fig. 1. Activity Diagram to GALS Systems Behavior Verification

- modeler activities: (1) modeling GALS system through IOPT nets; (2) splitting IOPT nets to obtain an IOPT net for each component of the GALS system;
- verifier activities: (3) translation from IOPT net models to Maude models; (4) translation from Maude models to Maude concrete syntax language; (5) specification of system properties to be verified; and (6) properties verification.

3.1 Running Example

The following example will be used through the paper to present the proposed methodology steps. The example is a very simplified condominium alarm system, which is used to detect events and control alarms of buildings. If an event occurs in one of the buildings, their alarm along with their neighbor buildings must ring.

In this example there are three buildings in a row, the building "1" has the neighbor building "2", the building "2" has the neighbor buildings "1" and "3", and building "3" has neighbor building "2".

3.2 System Modeling

This example was modeled by an IOPT net model, and is presented in Fig. 2. As previously mentioned, the IOPT Petri net is a class of non-autonomous Petri nets that extends the Place-Transition (P/T) Petri net class with inputs and outputs. These can be input and outputs signals and also input and output events.

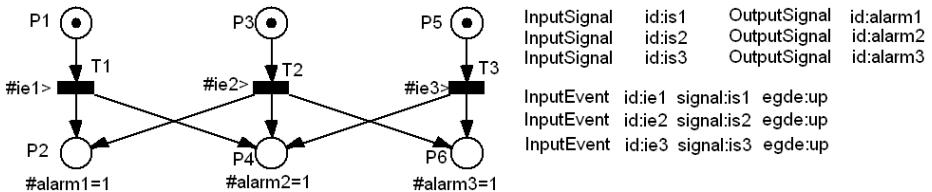


Fig. 2. IOPT model of an oversimplified condominium alarm system

Transition firing depends not only on the net marking, but also on the associated input events as well as the guard expression attached to the transition. Output expressions affecting output signals can be associated with places. When compared to Place-Transition nets, IOPT nets have other specific characteristics, as test arcs and priorities, which are described in [2].

The example has three input events ($ev1$, $ev2$, and $ev3$) associated with transitions governing the evolution of the IOPT net, and three output signals ($alarm1$, $alarm2$, and $alarm3$) that receive the value “1” when the corresponding place has one or more tokens.

3.3 Model Splitting

Considering that the system example will be implemented in a distributed way using three controllers (a controller in each building), the IOPT net model presented in Fig. 2, was divided into the three sub-models presented in Fig. 3, through the Net Splitting Operation.

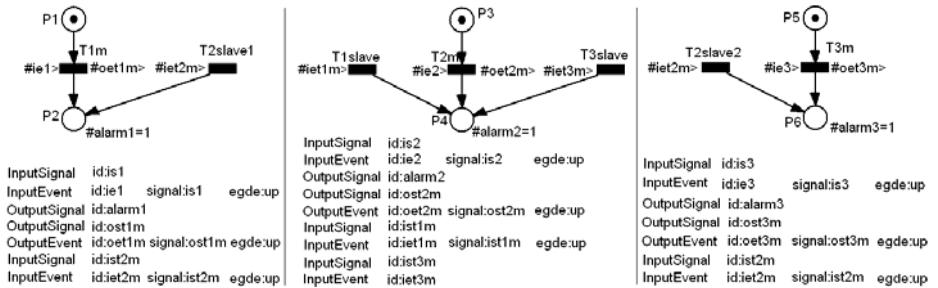


Fig. 3. IOPT sub-models resulting from the Net Splitting Operation

The first step of Net Splitting Operation is the definition of a valid cutting set, which finds a set of nodes with specific characteristics that will be used to divide the original net. The splitting was applied through the nodes $T1$, $T2$, and $T3$, generating the resulting sub-modules interconnected through the transitions $T1m/T1slave$, $T2m/T2slave1/T2slave2$, and $T3m/T3slave$, with associated output events (for instance $oet1m$) in the master transitions, that will be the input events (for instance $iet1m$) of slave transitions in other components. In this sense, the distributed execution model is composed by a set of parallel components communicating through a set of events.

4 Executable Semantics

GALS (Globally Asynchronous Locally Synchronous) systems are composed of several interacting components. Each component is synchronous, which means that its evolution is made at specific instants in time, controlled by a local clock. On the other hand, the global system is asynchronous. As there is no global clock synchronizing the components, each component is evolving at its own clock rate. The interaction between components is made sending messages through communication channels. In this sense, GALS systems have interleaving semantics.

IOPT nets were used in this work to model the whole GALS system, as well as GALS components. The firing of the transitions in one IOPT net (net evolution in one component) is done synchronously at specific instants in time (the synchronized paradigm), normally referred as tics or global clock; this means that, for that

component, between these instances, net marking will not change. The external clock or tic defines the moments in which enabled and ready transitions can fire. The enabled transition concept refers to the net marking dependency, as usual, while the ready transition concept is associated with the non-autonomous attributes evaluation. The IOPT nets have maximal step semantics, which means that all transitions that are enabled and ready at a specific instant in time will fire in that instant.

Fig. 4 presents a GALS system composed by three sub-models (components), each of them modeled with IOPT nets, representing the three components of the running example obtained through the net splitting operation (each of the clouds is associated with one sub-model of Fig. 3). Each sub-model will be potentially associated with a component running on an autonomous platform. The interaction between the various components is modeled through a set of events and accomplished through specific communication channels, for example direct connections, connections via asynchronous wrappers, NoC (network-on-chip), or any other type of networks, as common in distributed systems.

From the IOPT net model viewpoint, the border of each of these components is a set of nodes composed only by transitions. However, as far as the synchronous paradigm can not be applied to the whole system, the events used to assure the communication between components were replaced by places, modeling the separation of time instants associated with the firing of a master transition (emission of the output event from one component) and the firing of a slave transition (reception of the input event by the other component).

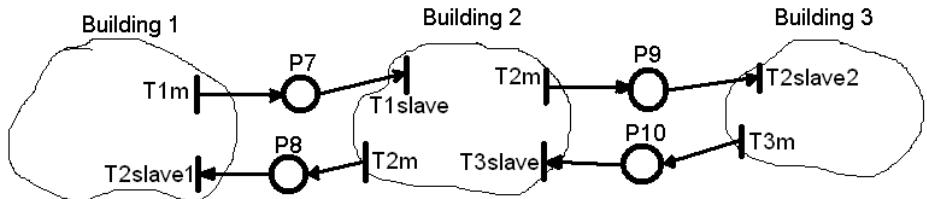


Fig. 4. IOPT nets modeling a GALS system

Each component is an IOPT net model with a maximal step execution, but the evolution tics of one component is different from the evolution tics of the other components, supporting the global asynchrony. In this sense, between components there is an interleaving execution semantics, while each component is governed by a maximal step execution semantics (each component is in a distinct execution temporal domain).

5 From IOPT Models of GALS Systems to Rewriting Logic Objects in Maude

5.1 Maude Language

The Maude language is a declarative language [4]. The basic programming statements are *equations* and *rules*, with simple rewriting semantics. *Rules* can be applied

concurrently, which means that *System modules* can be highly concurrent and non deterministic. In the Maude language it is possible to support objects and distributed objects interactions with rewrite *rules*. Objects interactions can be made through messages. Maude modules can be used: (1) as programs; (2) as executable specification; and (3) as models, that can be verified. In this work these modules will be used as models in which systems properties will be verified.

5.2 MDA Transformations

The MDA approach is used to make the transformation from IOPT net models to the concrete syntax of Maude language. Two transformations were made: (1) model-to-model transformation using the IOPT net metamodel proposed in [9] and Maude metamodel, and (2) model-to-text transformation to obtain Maude code. Model-to-model transformations were achieved using ATL transformation language and model-to-text transformations were made using MOFScript (a tool for model to text transformation).

5.3 GALS System Representation in Maude Language

GALS components modeled by IOPT nets are translated into Maude concurrent objects that interact through asynchronous messages. These messages represent the interleaving semantics of the *Globally Asynchronous* part of the GALS systems.

Maude code of GALS components obtained through the MDA transformation from IOPT net models have interleaving semantics (which is the naturally Maude semantics), although IOPT nets components have a maximal step semantics. This means that the behavior of this Maude code: (1) has exactly the same behavior of the IOPT net model, if and only if, in the each component IOPT net model at most one transition fires at each execution cycle, or (2) has a consistent behavior with the IOPT net model when, the change of, firing at most one transition per execution cycle over several execution cycles, rather than, firing several transitions in just one execution cycle, do not change the GALS components requirements/properties. In the running example, if the transitions $T1m$ and $T2slave1$ fires in the same execution cycle or if they fire in two consecutive execution cycles, the system requirements remain unchanged.

The generated Maude code for the running example is composed of 2 modules: *PETRI_NET_GALS* and *PETRI_NET_GALS_RULES*, an excerpt of it is presented below. Maude notation is presented in Maude manual in [4].

PETRI_NET_GALS module has the structure of the IOPT nets, in line 2 is made the inclusion of the *CONFIGURATION* module, which declares sorts representing concepts of objects, messages, and configurations that will be needed to represent the three IOPT nets, and the communication between them (represented by $P7$, $P8$, $P9$, and $P10$). In line 16 the three IOPT nets class identifiers are defined.

PETRI_NET_GALS_RULES module contains the transition rules, in line 31 is presented the rule for transition $T1m$, which removes one token from place $P1$ and creates one token in $P2$, and one token in $P7$, that represents a message going from component 1 (building 1) to component 2 (building 2).

```

1  mod PETRI_NET_GALS is
2    including CONFIGURATION .
3    sorts LocalTokens GlobalTokens Marking IOPT .
4    ops P1 P2 P3 P4 P5 P6      : -> LocalTokens .
5    ops P7 P8 P9 P10     : -> GlobalTokens .
6    ops Petri1 Petri2 Petri3 : -> Cid [ctor] .
7  endm
8
9  mod PETRI_NET_GALS_RULES is
10   protecting PETRI_NET_GALS .
11   protecting META-LEVEL .
12   var O1 O2 O3           : Object .
13   vars AP7 AP8 AP9 AP10   : GlobalTokens .
14   var petri              : Oid .
15   vars AP1 AP2 AP3 AP4 AP5 AP6 : LocalTokens .
16   var S                  : String .
17   rl [T1m] : < petri : Petri1 | m (P1 AP1,      AP2) >,
18          O2, O3, ( AP7, AP8, AP9, AP10), S =>
19          < petri : Petri1 | m (      AP1, P2 AP2) >,
20          O2, O3, (P7 AP7, AP8, AP9, AP10), "T1m" .
21
22 endm

```

5.4 Verification

As described in Section 3 about the example: (1) if an event occurs in building "1", the alarms of buildings "1" and "2" should begin to ring; (2) if an event occurs in building "2", the alarms of buildings "1", "2", and "3" should begin to ring; and (3) if an event occurs in building "3", the alarms of buildings "2" and "3" should begin to ring. These are the 3 properties that should be verified. Alarm of building "1" rings if the marking of place P1 is equal or greater than 1, and so on.

The generated Maude code, presented in section 5.3, was used in the Maude system to verify the three mentioned properties. To verify property one, it was checked all the possible final states of the system after event "1" occurs. To do this the associated state space containing all reachable states was generated and analyzed. To generate the state space in Maude, the search command (*search <petri1:Oid : Petri1 | m (P1, empty) >, <petri2:Oid : Petri2 | m (P3, empty) >, <petri3:Oid : Petri3 | m (P5, empty) >, (none, none, none, none), "" =>! Any:Net .*) was used in the code presented in section 5.3, and to show it the command (*show search graph .*) was used. It was verified that all possible final states of the system after event "1" occurrence are $(P2=1, P3=1, P4=1, P5=1)$, $(P2=2, P4=2, P5=1, P6=1)$, $(P2=1, P3=1, P4=2, P6=1)$ or $(P2=2, P4=3, P6=2)$. Analyzing them, it can be concluded that the places $P2$ and $P4$ have always one or more tokens, which means that alarms of buildings "1" and "2" ring in this situation, and can be concluded that property one is successfully verified. Properties two and three were also successfully verified.

Due to space limitations, it is not possible to present the complete state space, even for such simple example in order to make evident the referred verified properties. Instead, a simplified system composed by buildings 1 and 2 is considered. Fig. 5 presents the partial state space of this simplified system, considering the sub-models of Fig. 3 and Fig. 4 associated with buildings 1 and 2, which means the models with the transitions $T1m$, $T1slave$, $T2m$, and $T2slave1$, and places $P1$, $P2$, $P3$, $P4$, $P7$, and $P8$. This partial state space has 9 states, presented in Fig. 5, while the full state space

has 45 states. Each node of the state space presented in Fig. 5 shows the marking of the relevant places, where the four nodes with $P7$ and $P8$ holding no tokens are the ones observed in the initial model of Fig. 2.

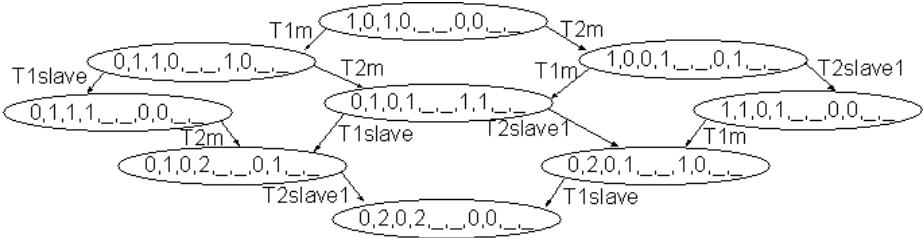


Fig. 5. Partial state space of the example

6 Related Work

Since most of the embedded systems circuitry is made of synchronous circuits, they became the starting point in the development of GALS systems. With this in mind, there are several works proposing architectures, property verification, implementations, and prototyping for GALS systems.

Some authors propose a verification approach for GALS systems (e.g. [10]), taking as starting point the description of the systems' behavior using textual languages, instead of graphical-based descriptions like the ones being proposed in this paper. Other authors use Petri nets to represent GALS systems behavior and to verify its properties (e.g. [11]); however, the methodology does not cover the entire GALS systems development flow, that starts with Petri net models. In [1], a full development flow for embedded systems was proposed through automatic code generation, from Petri net models, but without attempting to answer the specific questions of GALS systems.

To the best of our knowledge, no works addresses the complete development flow (modeling, simulation, verification, and implementation) of GALS systems through automatic code generation based on non-elementary Petri nets.

7 Conclusions

The methodology for specification and verification of GALS systems using IOPT nets as modeling formalism was shown to be adequate in the testing phase of this work. In all the validation examples it was possible to model distributed execution of embedded systems as a GALS system using IOPT nets, and to verify systems properties with Maude. Maude code was always automatically generated from IOPT nets through model-to-model and model-to-text transformations.

The main conclusion is that the proposed development methodology has several advantages compared to a development methodology that does not use models. It also takes advantage from the usage of Petri nets as the underlying model of computation,

namely (1) the model clearly describes the system's behavior; (2) it is possible to start modeling the system with one centralized model, which is then partitioned into a set of modules, the components of the GALS system; and (3) the system verification and implementation codes are automatically generated from the model, which decreases the development time and the potential errors of manual code generation.

Acknowledgments. This work is supported by the cooperation project funded by Portuguese FCT through the project ref. 4.4.1.00-CAPES, and by Brazilian CAPES through the project ref. 236/09. The first author work is supported by a Portuguese FCT (Fundação para a Ciência e a Tecnologia) grant, ref. SFRH/BD/62171/2009.

References

1. Gomes, L., Barros, J.P., Costa, A., Pais, R., Moutinho, F.: Formal Methods for Embedded Systems Co-design: the FORDESIGN Project. In: Proceedings of Workshop Reconfigurable Communication-centric Systems-on-Chip, ReCoSoC (2005)
2. Gomes, L., Barros, J., Costa, A., Nunes, R.: The Input-Output Place-Transition Petri Net Class and Associated Tools. In: Proceedings of the 5th IEEE International Conference on Industrial Informatics (INDIN 2007), Vienna, Austria (2007)
3. Costa, A., Gomes, L.: Petri net partitioning using net splitting operation. In: 7th IEEE International Conference on Industrial Informatics (INDIN 2009), Cardiff, UK (2009), <http://dx.doi.org/10.1109/INDIN.2009.5195804>
4. Clavel, M., Durán, F., Eker, S., Lincoln, P., Martí-Oliet, N., Meseguer, J., Talcott, C.: Maude Manual (Version 2.5), <http://maude.cs.uiuc.edu/maude2-manual/edn>
5. OMG-MDA, Omg mda guide version 1.0.1. formal doc.: (June -03- 2001), <http://www.omg.org/cgi-bin/doc?omg/03-06-01> (accessed January, 2010)
6. PNML2C: PNML2C - A translator from PNML to C, <http://www.uninova.pt/fordesign/PNML2C.htm> (accessed March 30, 2010)
7. PNML2VHDL: PNML2VHDL - A translator from PNML to VHDL, <http://www.uninova.pt/fordesign/PNML2VHDL.htm> (accessed March 30, 2010)
8. Billington, J., Christensen, S., van Hee, K.M., Kindler, E., Kummer, O., Petrucci, L., Post, R., Stehno, C., Weber, M.: The Petri Net Markup Language: Concepts, Technology, and Tools. In: van der Aalst, W.M.P., Best, E. (eds.) ICATPN 2003. LNCS, vol. 2679, pp. 483–505. Springer, Heidelberg (2003)
9. Moutinho, F., Gomes, L., Ramalho, F., Figueiredo, J., Barros, J., Barbosa, P., Pais, R., Costa, A.: Ecore Representation for Extending PNML for Input-Output Place-Transition Nets. In: 36th Annual Conference of the IEEE Industrial Electronics Society, IECON 2010, Phoenix, AZ, USA, November 7-10 (2010)
10. Doucet, F., Menarini, M., Kruger, I., Gupta, R.: A Verification Approach for GALS Integration of Synchronous Components. (2005), <http://www.irisa.fr/prive/talpin/papers/fmgals05a.pdf> (accessed July 25, 2010)
11. Dasgupta, S., Yakovlev, A.: Modeling and Performance Analysis of GALS Architectures. In: International Symposium on System-on-Chip 2006, Tampere, Finland (2006)

Automatic Generation of Run-Time Monitoring Capabilities to Petri Nets Based Controllers with Graphical User Interfaces

Fernando Pereira^{1,2}, Luis Gomes^{1,3}, and Filipe Moutinho^{1,3}

¹ FCT/UNL Universidade Nova de Lisboa

² ISEL Instituto Superior de Engenharia de Lisboa

³ UNINOVA, Portugal

fjp@deea.isel.ipl.pt, lugo@fct.unl.pt, fcm@uninova.pt

Abstract. The growing processing power available in FPGAs and other embedded platforms, associated with the ability to generate high resolution images and interface with pointing devices, opened the possibility to create devices with sophisticated user interfaces. This paper presents an innovative tool to automatically generate debug, diagnostic and monitoring graphical interfaces to be integrated in embedded systems designed using Petri net based controllers. Devices powered with the new debug and diagnostic interfaces benefit from lower maintenance costs and simplified failure diagnostic capabilities, leading to longer product life cycles with the corresponding environmental and sustainability gains. To demonstrate the validity of the tools proposed, the paper presents an application example for a Car Parking controller, including results on a working prototype.

Keywords: Embedded Systems, Petri nets, Design automation, Modeling, Graphical User Interfaces.

1 Introduction

Graphical debug and monitoring tools have always played a very important role in the development of embedded and automation systems. Due to the lack of resources and processing power available in hardware used to deploy these solutions, the tools have traditionally relied on software running on external personal computers.

However, the growing adoption of reconfigurable hardware platforms (ex. FPGAs) in embedded and industrial automation solutions brought increased processing power associated with the capability to generate high resolution images and interface with pointing devices, as mice and touch-screens, with no significant additional cost.

This paper presents a tool framework to the automatic generation of graphical debug, diagnostic and monitoring interfaces directly in FPGA hardware. This approach has many advantages over traditional solutions because the tools can be used after the development, test and validation phase terminates, to perform maintenance tasks and help diagnose mechanical and electrical faults.

The new tool takes advantage and extends a previous framework [1] containing design and modeling tools based on IOPT (Input-Output Place-Transition) Petri nets

[2], simulation and automatic code generation tools for micro-controllers or FPGAs, plus an Animator tool to produce Graphical User Interfaces associated with the IOPT model execution [3].

The proposed solution analyzes a PNML file [4] generated by the referred tool chain describing an embedded system controller and automatically creates a set of XML files to the Animator tool, containing a debug and monitoring animation screen. This screen contains a graphical image of the IOPT Petri net model and a set of animation rules to display the status of the model in real time, including net marking, transition status and input and output signals. The system designer can later integrate this animation screen in the final application user interface.

2 Contribution to Technological Innovation and Sustainability

The main innovation presented in this paper is the capability to automatically generate debug and monitoring graphical interfaces for embedded systems, with zero additional design effort and negligible cost. The complete tool chain can generate full embedded system controllers for FPGAs, including the controller and an animated GUI with a debug and monitoring interface, without writing a single line of code.

Adding graphical debug and monitoring interfaces to embedded devices can have an enormous environmental impact and greatly contribute for sustainability. To better understand those impacts, embedded devices should be separated into two classes: industrial automation systems and end-user appliances.

Industrial automation systems generally have high availability requirements because downtime in one system can stall entire production lines, causing effective downtime over entire production plants, with the consequent delivery delays and high labour loss. In light of this problem, the performance of maintenance and technical assistance services is regarded with special importance.

Technical assistance interventions are generally characterized by a typical pattern: whenever an assistance call is received by an equipment supplier, a technical team is immediately scheduled to visit the customer's site and diagnose the problem, returning home to fetch the required parts, followed by a second visit to implement a solution.

Embedding diagnostic and monitoring capabilities in the final systems can effectively break this pattern, as machine operators and the factory's maintenance engineers, with the help of the vendor's remote assistance, have the means to diagnose problems and identify damaged parts, reducing the number of travels to just one. Factory's maintenance engineers can even receive training to use auto-diagnostic systems to solve most problems and replacement parts can be sent using express mail services, avoiding the need to send technicians altogether. This solution results in faster repair times, minimized down-times and equipment suppliers can operate with smaller technical assistance vehicle fleets, reducing energy consumption and contributing for sustainability.

Another indirect result is the reduction of redundant production units: to minimize downtime, industrial facilities generally purchase spare units of the most sensitive machinery. The number of redundant units is calculated according to past failure statistics (MTBF) and average repair time. Lower average repair times enable the reduction of spare units, contributing even more to sustainability.

Embedded systems present in end-user appliances can also benefit from internal diagnostic and monitoring interfaces. In this class of systems, a high percentage of technical assistance incidents is related to improper user operation and bad device configuration, as users did not receive adequate training, resulting in unnecessary travels to support centers. The addition of internal diagnostic graphical interfaces provide an effective way to simplify the communication between end-users and help-desk staff, allowing to solve most problems remotely.

When devices suffer from real defects, internal debug and diagnostic interfaces can provide the same gains experienced by industrial systems: end-users and help-desk staff can cooperate, diagnosing failures remotely and making possible to send a technician with all the necessary parts to quickly solve the problem.

Due to the lack of capability to diagnose and solve problems in a single visit, most brand-name manufacturers strategy consists in immediately replacing systems covered by warranty with new units. This approach poses many environmental hazards, as the consumables present in the old systems are simply disposed as garbage and the old systems cannot be resold as new after repair, being often also disposed.

When appliances malfunction after warranties expire, there is a common perception that they are not worth repair, due to the low cost of new units, high transportation costs and high technical-center fees. This happens even when faults are caused by trivial problems as a loosen screw, a melted fuse or a wrong EPROM configuration.

The addition of integrated debug and diagnostic interfaces can help users detect most problems and repair the most trivial ones or use the service of local repair shops. These shops used to be very popular several decades ago, but the advent of ever increasing complex electronic devices, requiring the use of specialized diagnostic equipment, turned the repair of sophisticated electronic devices almost impossible. The addition of integrated diagnostic and debug interfaces can revive local repair shops and allow end users repair trivial problems, largely increasing the useful life cycle of consumer devices, minimizing the creation of hazardous garbage and contributing to great environmental and sustainability gains.

3 Comparison with Present Solutions

Debug and monitoring interfaces present in embedded systems development tools generally run on external computers connected to the physical embedded systems using special purpose data cables. This is the usual method employed in industrial programmable logic controllers. However, an industrial facility generally contains many systems from multiple vendors and it is not always possible to hold the development tools for all of them.

Even the development tools may not be enough to run diagnostics because the tools often require access to the real model files used during system development. However, equipment suppliers may deny access to the development model files to hide implementation secrets and to prevent unauthorized changes that may compromise safety and regulation compliance, leading to possible legal problems.

On the contrary, the diagnostic interfaces produced by the new tools are available to end users without requiring any dedicated hardware or software, yet do not allow

system changes. To prevent access to implementation secrets, vendors can choose to install a simplified debug model, providing enough information to diagnose failures, but hiding sensitive information.

Most current end-user appliances include some degree of self-test and diagnostic utilities. For example, some printers have self-test pages and many devices generate error codes presented as display messages or blinking LED counts.

However, the self-test routines must be specifically programmed during the development phase and only detect typical failures predicted by system developers. On the contrary, the new tools do not require programming and automatically append a debug and diagnostic interface to the final system, helping diagnose all types of failures, including those not originally foreseen.

More importantly, the traditional diagnostic routines do not work when a device reaches a deadlock situation. To perform diagnostics the device must be restarted, failing to identify transient error conditions. In alternative, the new diagnostic interfaces run in parallel with the system controller and can be recalled at any time, independently of the main controller status and without causing any state changes, thus allowing the detection of deadlock and transient faults.

Finally, the error codes generated by current devices often have no value to the end users because the meaning of the codes is only available to manufacturer's technical staff. On the opposite, the new debug and diagnostic interfaces provide an intuitive animated graphical user interface, displaying the state of the system in real time as a Petri net model, which is the underlying modeling formalism.

4 Related Work

Embedded systems design based on Petri net models has been the subject of many research publications, ranging from Low level nets [2][5] to High level colored nets [6][7][8].

Most development frameworks based on Petri net models include debug and visualization interfaces [8][9], to exhibit the system state, but these tools generally work only during simulations running on personal computers and have not been ported to physical embedded devices.

Some Petri net frameworks include tools to create interactive animations [8][10]. For instance, the Colored Petri net Tools, a very successful modeling tool-chain, contains animation design tools [11][12] to enhance user-friendliness and simplify communication with persons with no knowledge about Petri net formalisms.

Other authors have worked on automatic code generation from Petri net models [13][14][15], to allow the rapid development of embedded applications. These tools automatically generate software source code and low level hardware descriptions implementing the behavior described by the model.

The contribution presented in this paper is based on a previous work [1][3] combining automatic software and hardware co-generation with the capability to design animated graphical user interfaces for embedded systems. The animations are automatically generated from the original Petri net model and a set of rules associating the visibility and position of graphical objects, varying according to the system state evolution.

Using the previous tools, a designer could rapidly create embedded applications with sophisticated graphical user interfaces, without writing code and without needing deep understanding about software and hardware design, thus bringing embedded systems design to a much broader audience.

In fact, the previous generation of tools already included the capability to implement debug and diagnostic interfaces, but those interfaces had to be designed by human operators, drawing a background image and defining a large set of rules to display the system state, one at a time. However, this operation was repetitive, error prone and time consuming, specially with large models. In contrast, the new tool generates the desired results without any human intervention.

5 Development Flow

The tools introduced in this paper are based on a development flow presented in [16], beginning with the analysis of system behavior through the identification of patterns of use, leading to the creation of UML Use-Case diagrams. The captured use-cases will then be modeled using IOPT Petri nets [2], with the help of the Snoopy IOPT net editor. After a first version of the controller model is finished, the editor will generate a PNML [4] file, with the corresponding XML representation. This PNML file will be the base for all subsequent development steps.

IOPT nets are a non-autonomous Petri net class inheriting the characteristics from the well known Place-Transition nets [17], with the capability to associate input and output signals and events to model elements, enabling the deterministic specification of the interaction between models and the external environment. IOPTs also benefit from maximal step semantics, meaning that all autonomously enabled transitions will immediately fire as soon as the associated guard conditions and events are active. To enable automatic conflict resolution, IOPT nets include transition priorities and other characteristics, like test arcs and arc weights that improve modeling capabilities.

Automatic code generation is performed through two applications, PNML2VHDL and PNML2C, that create system controllers based on VHDL hardware descriptions or “C” source code, according to the desired embedded target device.

Using the PNML file as input to an “Animator” tool, the developer can create a graphical user interface, consisting of multiple screens containing animated graphical objects. The PNML model is used as a base to define a set of rules describing the evolution of the animated graphical objects.

The animations created in the previous step can be presented on a personal computer during model simulations, to debug, validate and correct the designed controller model. Another tool – GUIGen4FPGA - will automatically generate VHDL code and EEPROM image files, to enable the execution on FPGAs.

To finally generate a running embedded application it is necessary to configure the physical hardware platform, through the association of model signal names and events to real FPGA pins, using a Xilinx UCF file.

The proposed development flow includes a simulation and animation step where the user can test and correct the designed model, returning to the first development stage whenever incorrect behavior is detected. However, the validity of the simulation phase largely depends on the ability to also model and simulate the physical environment surrounding the embedded controller, which can be a complex task. In those cases the validity of the models can only be checked on a physical prototype.

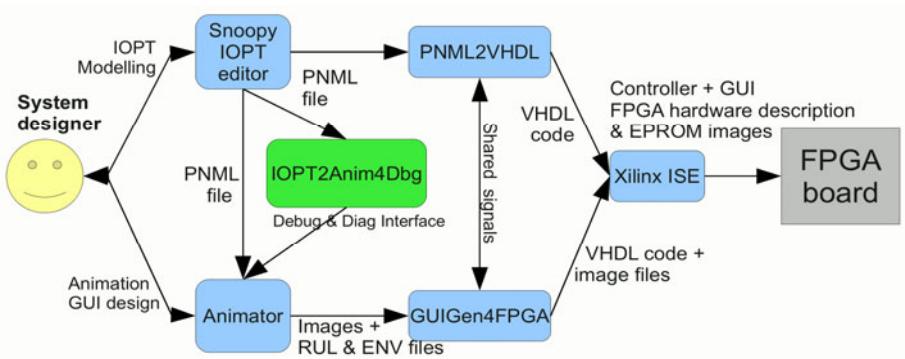


Fig. 1. Proposed development framework

One of the goals of the present work is to provide tools to help debug and identify mistakes during the prototype test phase. This way, the former development flow has been improved and a new tool “PNML2Anim4Dbg” was developed to automatically generate debug and diagnostic animation screens. Figure 1 displays a diagram describing the improved development tool chain.

6 Implementation

The tool introduced in this work, “PNML2Anim4Dbg”, receives input from a PNML file containing a IOPT Petri net model describing a system controller and automatically generates a set of XML files for the “Animator” tool.

As both the input and output files are encoded using XML, the XSLT (Extensible Style-sheet Transformation) [18] framework was selected to implement the new tool, due to the capability to automatically validate syntactic grammars using DTDs or Schema and XML pattern matching, XML tree navigation and query tools (XPATH).

The usage of XSL transformations applied to PNML files is as old as the PNML format itself, since the first documents introducing the PNML standard [4], already proposed XSL transformations as a tool to convert models between different Petri net classes.

As seen in figure 2, the PNML2Anim4Dbg creates 5 files: A SVG background image, a «rul» file containing all generated animation rules, an «env» file associating a list of BMP image files to shortcut names present in the animation rules, an «ov» file with a list of output values generated by the controller and a «pdc» file containing a list of input signals and the corresponding user interface methods.

Although other XSLT based PNML to SVG converters were available [19], a new IOPT2SVG transformation was created, to account with the non-autonomous nature of IOPTs, including input and output signal graphical representations, different test and normal arc representations, transition priorities, etc.

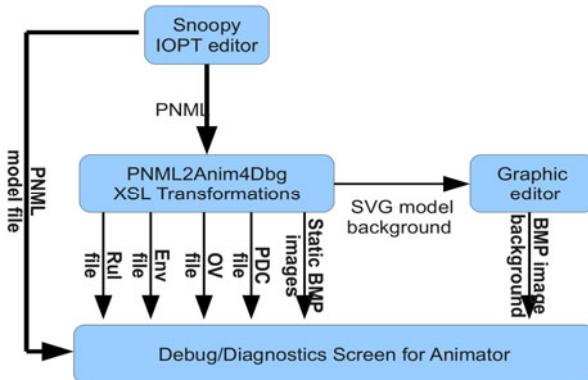


Fig. 2. PNML2Anim4Dbg data flow

The SVG background file contains an exact image of the IOPT model and can be edited and rearranged using most vector graphics editors, and finally converted to BMP format. During this phase it is possible to hide model comments and other superfluous information. Together with the SVG background image, a set of static image files is also appended to the animation project, containing pictures to display tokens inside places, signs to highlight the autonomously enabled transitions and LED signs to display active input and output signals.

The rules file contains a large set of rules to draw the correct number of tokens inside each place, to check if each transition is autonomously enabled and to check for active I/O signals. For more information about implementation details, the complete source code will be available online.

Complex hierarchical models, composed of several sub-net components, can be processed at different abstraction levels. Developers can choose simplified models showing only the top level components, full detailed models describing the entire system in a flat network, or create an hierarchy of multiple interface screens showing individual components. The tools do not require the entire model and can work with partial models, just requiring identifier consistency with the final controller system, maintaining the same place, transition and signal names.

7 Test and Validation

To test and validate the new tools, a car parking lot controller IOPT model (fig. 3) was processed using the proposed development flow to automatically generate a working prototype with a debug and diagnostic interface. This model is simple enough to eliminate the need for a detailed explanation, yet allows the demonstration of the proposed tools.

The hardware prototype was implemented using a Xilinx Spartan 3A 1800 Video Kit, including a Spartan 3A-DSP 1800 Starter Board, an Avnet EXP PS Video Module and a 1024x768 LCD panel. The starter board contains an FPGA and several memory devices, including a parallel flash EPROM to store images and a DDR2 RAM memory used to implement a video frame buffer.

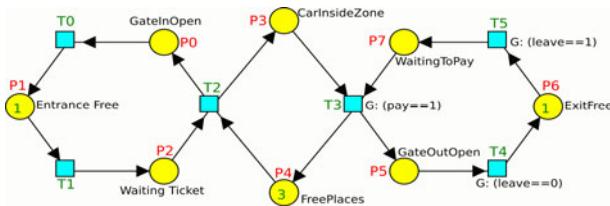


Fig. 3. Demonstration example IOPT model

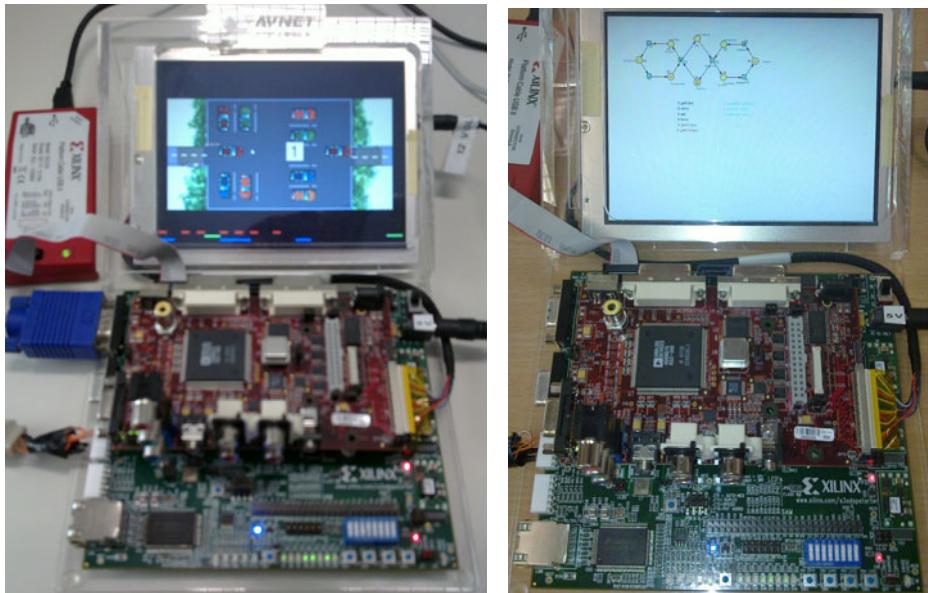


Fig. 4. Prototype photos – Application GUI on the left and debug interface on the right

Figure 4 displays photos of the prototype showing the Parking Lot animation screen and the corresponding debug and diagnostic screen. The car parking lot model contains one entrance, one exit, parking places for up to N vehicles, entry and exit barriers and inputs representing entry and payment sensors.

8 Conclusion and Future Work

An embedded system prototype was created and tested using the proposed tools, demonstrating the capability to rapidly generate embedded applications with sophisticated user interfaces and embedded debug, monitoring and diagnostic interfaces, using IOPT models and without the need to write any line of software code or design hardware components.

The new tools fully automate the task of diagnostics interface creation, generating a solution with minimal hardware requirements, just needing a few hundred Kbytes of EPROM space to store the debug interface images and some FPGA space to implement the animation rules. However, as most rules share code with the controller implementation, the VHDL logic optimizer tools will greatly simplify the generated hardware, removing duplicate signals. For example, both the controller and debug animation modules check if transitions are autonomously enabled and will share the same hardware. Image compression techniques, as simple as RLE encoding, can further reduce the total EPROM memory consumption to less than 50Kb per animation screen, with performance gains and no additional complexity.

Although the prototype was implemented using a standard FPGA development kit, it is possible to use the same tools to create very low cost production embedded controllers using dedicated PCB boards, requiring just one FPGA/ASIC chip, 2Mb of video RAM memory, one EEPROM, voltage regulators, interface logic and connectors, competing with the equivalent micro-controller based solutions.

As a result, in the near future will be possible to add debug and diagnostic interfaces to many embedded devices with no additional labour cost and irrelevant hardware cost increments, turning all the environmental and sustainability gains described in chapter 2 into a reality.

Future work include image the possibility to add pause and step-by-step execution capabilities to the generated interfaces. While the implementation of step-by-step execution mechanisms during simulation and software execution is trivial, it can pose technical challenges for hardware implementations, specially on asynchronous systems. Pausing and step-by-step execution on real hardware devices can create additional difficulties because certain real-time functions cannot be safely interrupted without the risk of causing permanent damages to external hardware and mechanical components. To solve this problem, additional work must be carried on.

Acknowledgment. The third author work is supported by a Portuguese FCT grant ref. SFRH/BD /62171/2009.

References

- [1] Moutinho, F., Gomes, L.: From models to controllers integrating graphical animation in FPGA through automatic code generation. In: IEEE International Symposium on Industrial Electronics (ISIE 2009), Seoul Olympic Parktel, Seoul, Korea, July 5-8 (2009)
- [2] Gomes, L., Barros, J., Costa, A., Nunes, R.: The Input-Output Place-Transition Petri Net Class and Associated Tools. In: Proceedings of the 5th IEEE International Conference on Industrial Informatics (INDIN 2007), Vienna, Austria (July 2007)
- [3] Gomes, L., Lourenco, J.: Rapid prototyping of graphical user interfaces for Petri-net-based controllers. *IEEE Transactions on Industrial Electronics* 57, 1806–1813 (2010)
- [4] Billington, J., Christensen, S., van Hee, K.M., Kindler, E., Kummer, O., Petrucci, L., Post, R., Stehno, C., Weber, M.: The Petri Net Markup Language: Concepts, Technology, and Tools. In: van der Aalst, W.M.P., Best, E. (eds.) ICATPN 2003. LNCS, vol. 2679, pp. 483–505. Springer, Heidelberg (2003)
- [5] Coolahan, J., Roussopoulos, N.: Timing requirements for time-driven systems using augmented Petri nets. *IEEE Transactions on Software Engineering*, 603–616 (September 1983)

- [6] Esser, R.: An object oriented Petri net language for embedded system design. In: Proceedings of the 8th International Workshop on Software Technology and Engineering Practice (STEP 1997) (including CASE 1997), p. 216. IEEE Computer Society, Washington, DC (1997)
- [7] Chachkov, S., Buchs, D.: From an abstract object-oriented model to a ready-to-use embedded system controller. In: 12th International Workshop on Rapid System Prototyping, Monterey, CA, pp. 142–148 (June 2001)
- [8] Jensen, K.: Coloured Petri Nets. Basic Concepts, Analysis Methods and Practical Use. Basic Concepts, vol. 1. Springer, Berlin (1997)
- [9] Kummer, O., Wienberg, F., Duvigneau, M., Cabac, L.: Renew – User Guide. University of Hamburg, Department for Informatics, Theoretical Foundations Group, Release 2.2, August 28 (2009)
- [10] Ehrig, H., Ermel, C., Taentzer, G.: Simulation and animation of visual models of embedded systems. In: 7th International Workshop on Embedded Systems Modeling Technology, and Applications, pp. 11–20 (June 2006)
- [11] Westergaard, M., Lassen, K.B.: The Britney suite animation tool. In: Donatelli, S., Thiagarajan, P.S. (eds.) ICATPN 2006. LNCS, vol. 4024, pp. 431–440. Springer, Heidelberg (2006)
- [12] Jorgensen, J.B.: Addressing problem frame concerns via Coloured Petri nets and graphical animation. In: 2006 International Workshop on Advances and Applications of Problem Frames, pp. 49–58 (May 2006)
- [13] Chachkov, S., Buchs, D.: From an abstract object-oriented model to a ready-to-use embedded system controller. In: 12th International Workshop on Rapid System Prototyping, Monterey, CA, pp. 142–148 (June 2001)
- [14] Nascimento, P., Maciel, P., Lima, M., Santana, R., Filho, A.: A partial reconfigurable architecture for controllers based on Petri nets. In: 17th Symposium on Integrated Circuits and System Design, pp. 16–21 (September 2004)
- [15] Costa, A., Gomes, L., Barros, J.P., Oliveira, J., Reis, T.: Petri nets tools framework supporting FPGA-based controller implementations. In: 34th Annual Conference of IEEE Industrial Electronics, IECON 2008, pp. 2477–2482 (2008), doi:10.1109/IECON.2008
- [16] Gomes, L., Barros, J.P., Costa, A., Pais, R., Moutinho, F.: Towards usage of formal methods within embedded systems co-design. In: 10th IEEE Conference on Emerging Technologies and Factory Automation, ETFA 2005, September 19-22, vol. 1, pp. 4–284 (2005), doi:10.1109/ETFA.2005.1612535
- [17] Reisig, W.: Petri nets: An introduction. Springer, New York (1985)
- [18] Tidwell, D.: XSLT. O'Reilly, Sebastopol (2001) ISBN 978-0-596-00053-0
- [19] ISO/IEC JTC1/SC7 N3298, ISO/IEC (2005)

SysVeritas: A Framework for Verifying IOPT Nets and Execution Semantics within Embedded Systems Design

Paulo Barbosa¹, João Paulo Barros^{2,4}, Franklin Ramalho¹, Luís Gomes^{3,4}, Jorge Figueiredo¹, Filipe Moutinho^{3,4}, Anikó Costa^{3,4}, and André Aranha¹

¹ Universidade Federal de Campina Grande, Campina Grande, Brazil

{paulo, franklin, abrantes, andre}@dsc.ufcg.edu.br

² Instituto Politécnico de Beja, Escola Superior de Tecnologia e Gestão, Portugal

³ Universidade Nova de Lisboa, Lisboa, Portugal

{lugo, fcm, jpb, akc}@uninova.pt

⁴ UNINOVA, Portugal

Abstract. We present a rewriting logic based technique for defining the formal executable semantics of a non-autonomous Petri net class, named Input-Output Place/Transition nets (IOPT nets), designed for model-based embedded system's development, according to the MDA initiative. For this purpose, we provide model-to-model transformations from ecore IOPT models to a rewriting logic specification in Maude. The transformations are defined as semantic mappings based on the respective metamodels: the IOPT metamodel and the Maude metamodel. Also, we define model to-text transformations for the generation of the model execution code in the rewriting logic framework. Hence, we present a translational semantics composed by two components: (i) the denotational one, considering as semantic domains the operations, equations, and properties that specify the Petri net structure, signals, and events according to the commutative monoid view; and (ii) the operational one, that changes the interleaving semantics of Maude using rewriting rules specified at the Maude metalevel to provide a maximal step semantics for transitions with arcs, test arcs, and priorities. Additionally, this work gives architectural advices in order to compose new semantics specifications by simple component substitution. Due to its simulation and verification capabilities for control systems, the presented work was applied to a domotic project that intends to save energy in residential buildings.

Keywords: Embedded Systems, Petri Nets, Maude, Verification.

1 Introduction

It is well accepted that models offer one of the best choices to deal with the development of complex systems [1]. In particular, models improve the communication between developers and customers. However, much more is expected from a single model. For example, in the embedded systems domain, one has to specify the system in an unambiguously way and enable sophisticated system analysis. Due to this fact, and in order to increase consistency, most of the currently accepted model-based development techniques are based on formal models [2], these models are able to precisely represent the semantics of computation and of concurrency.

Several distinct modeling formalisms, supporting the model-based development attitude [3], have already proved their adequacy for embedded systems design. With Petri nets [4] as a system specification language we get the advantages of its strong mathematical definition and its well defined and precise semantics, enabling the support for simulation, state space generation, and model-checking techniques for verification purposes.

From another point of view, Petri nets are suitable for the definition of automatic model transformations, in order to obtain models at different levels of abstraction. For example, the FORDESIGN project [3] has obtained several interesting results by reusing the benefits from transformations involving models defined using a non-autonomous Petri nets class entitled Input-Output Place/Transition nets (IOPT nets) and the respective code generation for several languages and platforms. Currently, the MDA-Veritas initiative [5] has been proposed as one alternative in order to reuse the state of the art in model based development provided by the FORDESIGN project, shifting the focus to MDA (Model-Driven Architecture) [6] thus obtaining improvements in the verification of IOPT nets as formal models.

Concerning the aforementioned context, the expected results are highly dependent on the semantics of the IOPT models. Since these models are expected to be built as Platform Independent Models (PIMs), they should be properly transformed into Platform Specific Models (PSMs) for mapping to the corresponding code. However, this automatic generation must take into account the platforms architecture and definition. Thus, an important question arises: how to obtain the verification goals, given the number of existing platforms? Models need to make sense in the corresponding environment they are inserted. Thus, since each specific platform has its own concepts and execution semantics, models should be able to represent these characteristics.

In Section 2, we will discuss about the environmental impacts of deploying a semantic model for PSMs, avoiding the early use of electronic devices. We explore the previously defined semantics for IOPTs, briefly recapitulated in Section 3. In Section 4, an executable formal model upon which deploying environmental characteristics will be represented. Moreover, through a running example in Section 5, the analysis for static properties, synthesis of executable sequential implementations, automated distribution and dynamical behaviors following model checking techniques are presented as one of the main points of this work. The approach is fully supported by MDA standards and tools and takes advantage of the suitability of Maude, and its metalanguage capabilities, for reactive systems modeling and execution. Finally, Section 6 discusses our final view about this gap between the concepts available in a design language, in this case the IOPT nets, and those available in platforms, as well as the analysis and verification tools. We expect that designers should be able to work with the domain-specific abstractions such as signals and events, so the knowledge required to map them into a platform must be provided automatically.

2 Contributions to Sustainability

We expect that the greater the effort to tackle the question of representing platform-specific characteristics in verification models, the greater will be the obtained benefits. More specifically, the creation of formal models eases the simulation and verification at several levels, hence decreasing the necessity of dealing with hardware

devices at early stages, avoiding the use of additional electronic devices that would hardly be recycled, saving energy, increasing the reliability level, and reducing costs. Hence, our approach constitutes a technological innovation that contributes to sustainability. To that end, we have identified four distinct contributions: (i) the use of more reliable development flows, due to the consolidation of a formal MDA approach, (ii) savings in costs due to hardware, energy, specialized engineer views (for the software developer), more abstract specifications, and correctness, (iii) the use of specific executable semantics specifications available in several kinds of platforms and (iv) at the practical level, the facilitation of rapid prototyping for reliable control and domotic systems.

3 IOPT Nets and MDA-Veritas

Petri nets are a well-known set of formal languages with a common graphical representation, particularly suitable for the visual modeling of concurrent systems. The class of Input Output Place Transition nets extends the class of Place/Transition nets (P/T nets) with non-autonomous constructs allowing the modeling of controllers connected to the environment through signals and events. The respective IOPT semantics that interest us was already presented elsewhere [3]. Next, we briefly present the main characteristics of this semantics.

IOPT nets add several annotations to the P/T nets nodes and net modules. More specifically, transitions can have associated input and output events, as well as a guard that is a function of the input signal values. Additionally, each conflict is resolved through the addition of a priority annotation to each transition. Places have a bound annotation specifying the maximum number of tokens in each place, which can be of major relevance when automatic code generation is considered. They also have a conditional external action on output signals. The respective condition is a function of the place marking. All signals and events are defined at the net module level. Next we briefly present the IOPT nets semantics.

The IOPT nets have *maximal step semantics*: whenever a transition is enabled, and the associated external condition is true (the input event and the input signal guard are both true), the transition is fired. An IOPT net step is maximum when no additional transition can be fired without generating an effective conflict with some transition in the chosen maximal step. Therefore, we define a IOPT net step occurrence and the respective successor marking. The net evolves through the firing of successive maximal steps. The synchronized paradigm, used in this work, also implies that the net evolution is only possible at specific instants in time named *tics*. These are defined by an external *global clock*. An execution step is the period between two tics.

3.1 System Modeling

In [7] we have proposed a formal approach named *semantic equations* to extract formal models from the syntactic constructors, *i.e.* the metamodel. There, we have approached Petri nets models involved in model transformations aiming to ensure semantics preserving properties. Here, we are interested in reusing the *semantic equations* approach but for another purpose: to extract the state space as the semantic

representation for the IOPT nets formalism. The understanding of this state space in terms of algebraic structures over graphs has the potential to unify several views of semantics formalisms. This structure can be manipulated by formal methods tools employed in several sorts of analysis and satisfying some requirements that are hard to solve. As an example, several formal tools could handle the state space explosion problem, by providing an on-the-fly way of reasoning over occurrence graph structure of the Petri net model if it is seen as this algebraic structure.

As in [7], this ordinary graph has a set of nodes as a commutative monoid $S+$ generated by the set S of places and the empty marking as identity. The sum $+$ of transitions represents the parallel firing of transitions, and the sequential firing of transitions is denoted by the operation. By the guarantee of the closure property, we are representing computations through simple transitions.

The concept of *semantic domain* and *semantic equations* gives us a guarantee that the metamodel presented in [8] will provide models able to satisfy the desired executability requirement.

4 From IOPT Models to Algebraic Specifications

Fig. 1 shows the main flow for the use of this solution. It is inserted in a major flow of the MDA-Veritas solution. We have two roles: (i) the modeler that starts editing the model until, throughout MDA transformations, generating the Maude code; and (ii) the verifier, that employs the Maude model in concrete syntax until the final analysis of formulae. The formulae can be proved at the object level by using the Maude LTL model-checker through the definition of predicates.

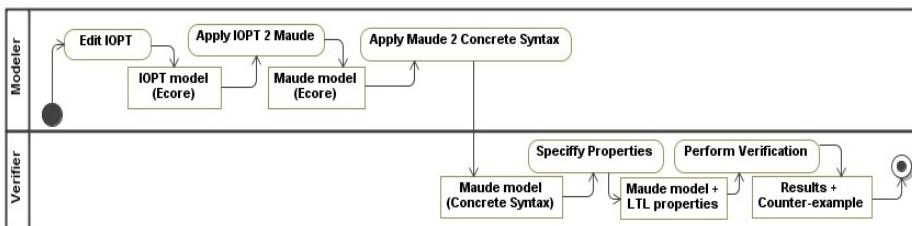
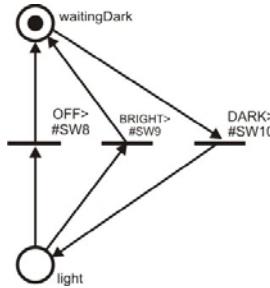


Fig. 1. Modeling and verification process

Fig. 2 presents a very simple example illustrating how to create and derive the *semantic model* for an IOPT. The model reuses the basic Petri nets graph structure. The following generated code has this graph representation plus the new features, such as events (OFF, BRIGHT, and DARK) and signals (SW8, SW9, and SW10). However, the most important features cannot be seen syntactically, because they are in the semantic domain. Examples of this are the specific execution semantics adopted for the model, the priority decision that decides what transition will fire in a conflict situation, or the occurrence of signals and events that affect the guards of the Petri net.

**Fig. 2.** Fragment of a domotics model

For example, the following Maude code represents the net of Fig. 2. It describes the existing sorts (line 2) for this algebraic representation, existing events (line 3), existing signals (line 4), and existing places (line 5). Other constructions responsible for the soundness of the definition will be discussed next. Finally, we have the representation of one transition (line 6), called turnLightOn, which has as precondition the DARK event, the SW10 signal, the waitingDark token, and produces no event (idle), no signal (noSignal), and one token light.

```

1 mod DOMOTICS-IOPT-NET-CONFIGURATION is
2 sorts Event EventSet Signal SignalSet PlaceMarking NetMarking IOPT .
3 ops OFF BRIGHT DARK : -> Event . op idle : -> Event .
4 ops SW8 SW9 SW10 : -> Signal . op noSignal : -> Signal .
5 ops waitingDark light : -> PlaceMarking . op empty : -> PlaceMarking .
...
6 rl [turnLightOn]:{DARK}+[SW10]+(waitngDark)=>{idle}+[noSignal]+(light) .
7 endm

```

The following code fragments represent a basic template for the translation of an IOPT model to a rewriting logic specification in Maude. We use regular grammar constructors for specifying the possible number of elements in a model. Line 1 is the declaration of the module able to represent the IOPT net. Line 2 defines the existing sorts for this specification. From line 3 up to 5 we have the declaration of the names for sorts Event, Signal, and PlaceMarking respectively, and the corresponding identity operations, *i.e.* idle, noSignal and empty. Finally, line 6 is the basic operation of combination of PlaceMarkings that follows the principles of commutative monoid Petri nets representation with the corresponding properties.

```

1 mod name-IOPT-NET-CONFIGURATION is
2 sorts Event EventSet Signal SignalSet PlaceMarking NetMarking IOPT .
3 ops (Event_names) : -> Event . op idle : -> Event .
4 ops (Signal_names)* : -> Signal . op noSignal : -> Signal .
5 ops (PlaceMarking_names)* : -> PlaceMarking . op empty:-> PlaceMarking .
6 op __:PlaceMarking PlaceMarking ->PlaceMarking [assoc comm id: empty] .

```

Thus, the Event, Signal and PlaceMarking elements are composed, in lines 7 up to 9, in EventSet, SignalSet and NetMarking respectively. The composition of these major structures is called a IOPT type at line 11.

```

7 op {(_)*} : (Event)* -> EventSet [ctor] .
8 op [(_)*] : (Signal)* -> SignalSet [ctor] .
9 op ((())* : (Place)* -> NetMarking .
10 op noState : -> [IOPT] .
11 op _+_+_- : EventSet SignalSet NetMarking -> IOPT .

```

Finally, the transitions are represented as rewrite rules from one IOPT configuration (line 12) to another IOPT configuration (line 13).

```

(
12 rl [Rule_name]:{((Event_names)*)+((Signal_names)*)+((Place_names)*)
13 => {((Event_names)*)+((Signal_names)*)+((Place_names)*)} .
)*

```

4.1 Maximal Step Semantics

In order to give a semantic representation of Petri nets having the maximal step semantics in a translational way, we need to change the Maude's original interleaving semantics. The main structure of the modules produced in this activity is depicted in Fig. 3. Starting from an IOPT_SYSTEM_CONFIGURATION, our choice relies on the fact that the Maude system contains a predefined module called CONFIGURATION for defining a denotational semantics and a META-LEVEL for redefining the operational semantics. The components from the denotational view were presented in the previous section, through the definition of an IOPT_NET_CONFIGURATION. Here, we focus on the operational view. A the META-LEVEL, we can find operators that represent terms and modules. From this level, the module also supplies efficient descent operations reducing the computations from the metalevel to the object level. Thus, operations such as *metaApply*, that matches the term with lefthand side of the rule, apply the rule at the top of the term and returns the metarepresentation of the term, can be used in order to take the application of all enabled transitions at a single step. Inheriting from the META-LEVEL, we defined a module META-PETRI-NET able to represent the main structural operations and rules from a Petri net. Finally, the MAXIMAL_STEP redefines the execution semantics of the Maude system for this domain, according to our specification. An excerpt of this module is shown in the following code fragment.

```

1 rl [maximal-step] : T:Term =>
2 maxStep(T:Term, applicableRules(T:Term, rules, module)) .
...
3 eq maxStep(T:Term, (rl X:Term => Y:Term [label(Q:Qid)] .) RS:RuleSet) =
4 maxStep(getTerm(metaApply(module, T:Term, Q:Qid, none, 0)), RS:RuleSet)
5 eq maxStep(T:Term, none) = T:Term .

```

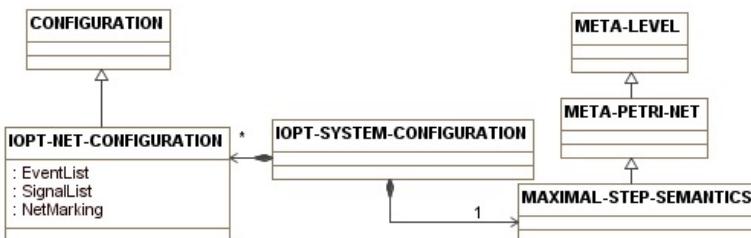


Fig. 3. The modules infrastructure

Lines 1 and 2 of this excerpt present the basic definition of the maximal-step rule for IOPT nets. A Term t is rewritten of the application of all applicable rules for the given module. Lines 3 and 4 detail the match of a given rule if applicable and the call of the built-in operation metaApply from the Maude META-LEVEL module. Line 5 represents the can in which all applicable rules were already applied, returning the current term.

5 Example

We have applied this formal specification in different scenarios. In order to illustrate it in practice, we show in Fig. 4 an application under development that supports domotic control systems. For simplification purposes, we have chosen a model with sensors that controls: (i) the arrive and leaving of strange people; (ii) the detection fire and its disabling; (iii) the alert for an unexpected situation and its stop; (iv) the detection of darkness and turning on the lights; and (v) a central enabler that establishes the priorities of these independent actuators.

From this model, we can automatically extract a *translational semantic* representation able to be employed in the Maude system. The following excerpt illustrates the single transition ***disFireAlert*** in Fig. 4 generated automatically from the previously described model.

```

1 mod DOMOTICS-IOPT-NET is
...
2 r1 [disFireAlert] : {DISABLE} + [SW4] + (enabled firing) =>
3 {noEvent} + [noSignal] + (waitingForFire waitingToEnable) .
...

```

This means that the transition ***disFireAlert***, which represents disabling the alert of fire, has dependencies in its firing from the existence of the event DISABLE and the signal SW4 (representing the action of pressing the switch of number 4 in the board that contains a PIC microcontroller) and having tokens in the places enabled and firing. As result, because of the act of consuming, no event and no signal are available in the system, the matched tokens were removed and new tokens for the places ***waitingForFire*** and ***waitingToEnable*** were produced. This represents a Platform-Specific Model (PSM) according to the MDA view.

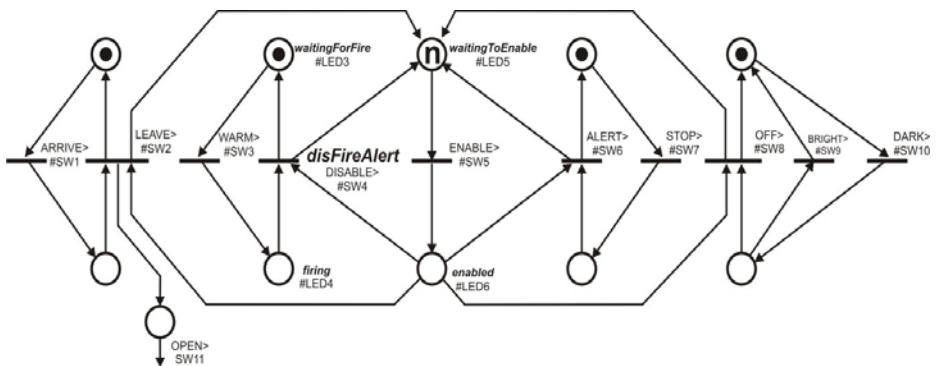


Fig. 4. Simplified domotics model

5.1 Simulation and Verification

The system can be simulated through common rewrite commands or generate the state space as an excerpt shown in Fig. 5. We developed a plugin to produce graphical visualizations of IOPT states after simulation and verification by using the GraphViz solution [9]. This represents the case where the system just allows the firing of one actuator that disables according to the priorities one enabled module from many (defined as $n=1$ in the place waitingToEnable) when sensors were fired concurrently. It shows that from the initial state, it is possible, through the maximal-step semantics, the activation of the events ARRIVE, WARM, ENABLE, STOP and DARK. From this state, excepting the luminosity module that is completely independent, only the ALERT event and its corresponding module, sensors and actuators are able to return to the initial state, depending from the ENABLE actuator.

The developers of the domotic model were interested in the verification of properties such as deadlock freeness, ensuring the correct application of the priorities and logical implications that given the firing of a sensor, the corresponding actuator will take the control and after solving will go back to the waiting state. These kind of primitive properties are automatically derived for the Maude LTL model-checker syntax from the initial system marking as the example that follows:

```
1 eq initial = {noEvent} + [noSignal] + (waitingForPresence
waitingForFire
2 waitingForDarkness waitingAlert waitingToEnable waitingToEnable)
...
3 search in DOMOTICS-IOPT-NET : upTerm(initial) =>! Any:Term .
```

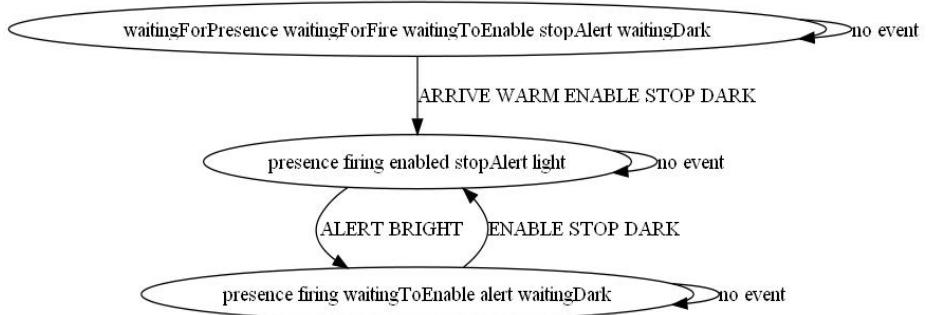


Fig. 5. Excerpt of the generated state space with the maximal step semantics

There, we have a search for a deadlocked state according to the maximal step semantics. The generated verification code starts by translating the initial marking specified in the.ecore IOPT model to the lines 1-2. This initial marking specifies that there are no events and signals and the initial tokens. Finally, line 3 represents the excerpt where the search Maude command is performed by translating the initial marking to its representation at the meta-level (upTerm) and try to find any term with no successors (through the command $=\Rightarrow!$). After successive refinements during the modeling phase, having several improvements, we get a running model with no deadlocks as show the following Maude output.

```
1 No solution .
2 states: 36 rewrites: 66 in 0ms cpu (0ms real) (~ rewrites/second)
```

6 Conclusions

For the embedded system's development, our contribution comes from the fact that safety and economic concerns require high levels of assurance that the designed model will work as expected in a physical environment. In this sense, the move to the physical implementation represents a huge gap of abstraction. Formal models enable us to solve this problem in an elegant fashion. The gap is filled with an artifact that is a representation of the designed model but can also represent logically most of conditions generated by the environment. Therefore we continue in the MDA lifecycle because this technique also represents the conceptual transformations of PIMs to PSMs according to the MDA view.

The present solution still has some limitations concerning the rigorous formalization of some execution semantics regarding the checking of semantics preservation in model transformations, establishing an equivalence relation between the models. More specifically, and although this does not affect the simulation and verification purposes initially established, the state space in the *maximal-step* semantics case cannot be fully explored for all partial subsets of events generated by the environment. As future work, we intend to extend the solution for more specific platforms, producing a semantic framework that will bring several benefits for the system's designer before producing a device from a chip layout or deploying code in an embedded platform.

Acknowledgment. This work is supported by the cooperation project funded by Portuguese FCT through the project ref. 4.4.1.00-CAPES, and by Brazilian CAPES through the project ref. 236/09.

References

1. Gomes, L., Fernandes, J.M.: Behavioral Modeling for Embedded Systems and Technologies. Information Science Reference (2009)
2. Sgroi, M., Lavagno, L., Sangiovanni-Vincentelli, A.: Formal models for embedded system design. IEEE Des. Test 17, 14–27 (2000)
3. Gomes, L., Barros, J.P., Costa, A., Pais, R., Moutinho, F.: Formal Methods for Embedded Systems Co design: the FORDESIGN Project. In: Proceedings of Workshop Reconfigurable Communication-centric Systems-on Chip, ReCoSoC 2005 (2005)
4. Girault, C., Rüdiger, V.: Petri Nets for Systems Engineering. In: XV, Hardcover, p. 607 (2003)
5. Barbosa, P., Costa, A., Ramalho, F., Figueiredo, J., Gomes, L., Junior, A.: Checking Semantics Equivalence of MDA Transformations in Concurrent Systems. Journal of Universal Computer Science (J.UCS) (to appear 2009), <http://www.jucs.org/jucs>
6. OMG: Model-Driven Architecture (2008), <http://www.omg.org/mda/>

7. Barbosa, P., Ramalho, F., Figueiredo, J., Costa, A., Gomes, L., Junior, A.: Semantic equations for formal models in the model-driven architecture. In: Camarinha-Matos, L.M., Pereira, P., Ribeiro, L. (eds.) DoCEIS 2010. IFIP Advances in Information and Communication Technology, vol. 314, pp. 251–260. Springer, Heidelberg (2010)
8. Moutinho, F., Gomes, L., Ramalho, F., Figueiredo, J., Barros, J.P., Barbosa, P., Pais, R., Costa, A.: Ecore Representation for Extending PNML for Input-Output Place-Transition Nets. In: 36th Annual Conference of the IEEE Industrial Electronics Society, IECON 2010, Glendale, AZ, November 7-10 (2010)
9. Ellson, J., Gansner, E., Koutsofios, L., North, S., Woodhull, G., Description, S., Technologies, L.: Graphviz. In: Open Source Graph Drawing Tools. LNCS, pp. 483–484. Springer, Heidelberg (2001)

Automatic Speech Recognition: An Improved Paradigm

Tudor-Sabin Topoleanu and Gheorghe Leonte Mogan

B-dul Eroilor, Nr. 29, 500036, Brasov, Romania

{tudor.topoleanu, mogan}@unitbv.ro

Abstract. In this paper we present a short survey of automatic speech recognition systems underlining the current achievements and capabilities of current day solutions as well as their inherent limitations and shortcomings. In response to which we propose an improved paradigm and algorithm for building an automatic speech recognition system that actively adapts its recognition model in an unsupervised fashion by listening to continuous human speech. The paradigm relies on creating a semi-autonomous system that samples continuous human speech in order to record phonetic units. Then processes those phoneme sized samples to identify the degree of similarity of each sample that will allow the detection of the same phoneme across many samples. After a sufficiently large database of samples has been gathered the system clusters the samples based on their degree of similarity, creating a different cluster for each phoneme. After that the system trains one neural network for each cluster using the samples in that cluster. After a few iterations of sampling, processing, clustering and training the system should contain a neural network detector for each phoneme unit of the spoken language that the system has been exposed to, and be able to use these detectors to recognize phonemes from live speech. Finally we provide the structure and algorithms for this novel automatic speech recognition paradigm.

Keywords: automatic speech recognition, natural language processing, probabilistic language acquisition, unsupervised learning of speech.

1 Introduction

Speech recognition is the process which transforms vocal sounds into the meaning of these sounds, turning spoken language into written language or symbolic knowledge, and it can be either human or automatic.

Human speech recognition turns spoken language into an internal symbolic representation in our minds, thus turning speech into meaning. The process of human speech recognition is based on sequentially recognizing phonetic units by taking advantage of multiple acoustic cues and then aligning them to obtain a word or sentence, this process happens in our subconscious mind without our constant attention [1].

Automatic speech recognition systems use the same principle of sequentially recognizing speech units from an audio signal based on recognition models that have been pre-trained to recognize these speech units, and then inferring the most probable word that is described by the succession of recognized speech units [2-3].

One feature that is not yet possible with the latter is the autonomous acquisition of the knowledge needed to recognize speech. Automatic systems are manually trained using databases of speech sounds [4], while humans are not born with the gift of speech recognition, instead they acquire it independently, progressively and autonomously [5].

Looking into the process of human language acquisition it becomes clear that it is a probabilistic endeavor [6]. Therefore it is possible to tackle this problem in a computational manner in order to program machines to acquire speech in an unsupervised manner [7-8]. However language acquisition in humans is an incremental process that starts with acquiring the capacity to recognize speech and then progressing to the process of language learning which relies on that former capacity [9]. Current research also suggests a strong link between perception and production of speech as these two processes constantly influence each other [10].

In this paper we propose an algorithm and structure for an automatic speech recognition system that allows semi-autonomous acquisition of speech recognition. The structure of the article is as follows: chapter two describes the contribution of our system to sustainability, chapter three is a short survey of automatic speech recognition systems, chapter four describes the structure algorithm and process that we propose for achieving semi-autonomous acquisition of speech recognition, chapter five addresses our current results and chapter six summarizes and details further work that will complete our research.

2 Contribution to Sustainability

The capacity to autonomously acquire, adapt and manage the database of speech samples needed to train neural networks for detecting phonetic units is the core innovation of our proposal. This ability gives our solution a higher level of autonomy and hence sustainability compared to current automatic speech recognition solutions.

This innovation gives the speech recognition system the capacity to self-maintain and also to adapt its database according to the inputs it receives, while also allowing the system to acquire its own samples of the language it will evolve to recognize. The purpose is to create a recognition system that needs little intervention from a human operator in order to be trained by being able to manage and train itself by processing its audio input.

Our intention is to create a system that mimics the human capability of acquiring speech and is therefore a self-sustaining software system that acts as a voice to text interface for other software systems. In our case the motivation for this research comes from creating an autonomous voice interface for mobile robots that will become a component module of a control architecture for mobile robots. Following this key requirement of self sustainability we set out to design a system that will be capable of acquiring the skill of speech recognition. However we do not think that our proposed solution will be useful only for mobile robot applications, we hope that it will find a use in other domains as well, for this reason we want to make it as easy as possible to exchange the recognition knowledge between instances of our system in order to allow other researchers to avoid, if so desired, the semi-autonomous acquisition phase.

3 A Short Survey of Automatic Speech Recognition

State of the art speech recognition systems can be split into a few categories: voice detection algorithms, user voice recognition, automatic speech recognition, emotion recognition and natural language processing.

Voice detection algorithms (VAD) simply detect when a recorded or live audio signal contains voice signals. One state of the art VAD system uses wavelet transforms in a wavelet filter bank for feature extraction from the input signal and then uses a Support Vector Machine (SVM) to train an optimized decision rule based on those extracted features [2, 11]. Another method that uses statistical models and machine learning for VAD employs generalized gamma distribution and learns from a speech database using minimum classification error (MCE) and SVM [12]. Another approach for robust VAD uses wavelet packet transform to analyze and extract transient components of speech and can extract speech activity from sources with a poor signal to noise ratio [13].

The current paradigm for automatic speech recognition consists of using either discriminative training models or generative training models [3]. Discriminative models based on Hidden Markov Models (HMMs) that are trained using a speech database and then used to recognize speech are a very important approach to realizing automatic speech recognition and could be extended to every corner of recognizer design [14]. Another discriminative approach to automatic speech recognition is based on neural networks which can be used for recognizing phonetic units, syllables or words [15] or the meaning of natural language [16].

The problem of active learning for speech recognition has been tackled before [17-18] however the approach is somewhat different since it relies on statistical processing and annotated corpus for processing real input data and minimizing the uncertainties from within it.

A current common trend is to create speech recognition systems that integrate multiple phonetic and acoustic feature extraction methods with language level modeling or language processing to reduce recognition errors and increase robustness of the system [19-22]. This is helpful because it provides multiple ways of detecting and eliminating recognition errors.

The common limitations and shortcomings of speech recognition systems is the requirement of using database of speech sounds for manually training the recognition models. The lack of support for less common languages, due to insufficient resources for compiling speech databases, and the high level of knowledge and technical skill required to train such recognition models and to create ASR systems.

4 An Improved Paradigm

Our proposed structure uses three levels for storing speech samples. The first level is a temporary buffer which has the function of storing all recorded speech samples until a limit has been reached. This limit is a parameter of our system (L1), once the limit is reached all samples from the buffer are analyzed and the useful ones are transferred to the second level of storage for further processing while the un-useful samples are removed, hence the buffer is now empty and ready for a new iteration.

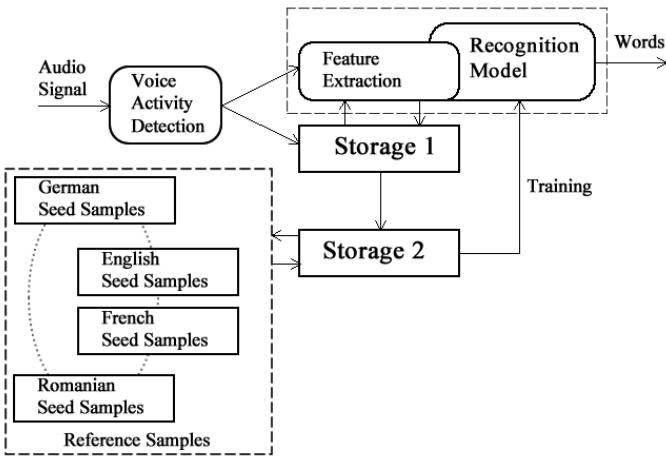


Fig. 1. The general structure of our system showcasing the main modules and data flow between them and the seed samples needed for clustering into phonetic clusters. As well as the separation between the recognition process, the acquisition process and the reference samples needed for clustering.

The second level of storage has the purpose of storing the samples selected from the temporary buffer and clustering them according to how similar they are. This level also has a limit of samples that can be stored, this is the second system parameter (L2), when the limit is reached clustering of samples is initiated at the end of each clustering process the last half of samples in each cluster are deleted. The selection of samples is made using a genetic algorithm that evaluates the fitness of each one and keeps only the best half of samples in each cluster.

The third and final level contains the neural networks that are trained with the samples in each cluster after the halving deletion. This level contains the models that are actually used to recognize phonetic units from live speech and therefore this level contains the recognition knowledge that can be transferred between instances of our system. This knowledge export/import feature is necessary since we consider that it would simplify the process of testing and evaluation before implementing a fully functional recognition system on a mobile robot, rather than testing and evaluating the system directly on the mobile robot.

Each neural network detector has the task of identifying a phonetic unit and is linked to a node in a HMM. In order to resolve the problem of the correct identification and clustering of phonetic units for each detector cluster the application will have to make the connection between phonetic unit and the equivalent written phoneme. In order to solve this problem we provide an “innate” set of samples representative of every phonetic unit of the language considered that the system contains from the start, these will be called reference or seed samples (these can be samples used in existing speech databases, or made especially for this task by requesting users to speak a pre-defined paragraph in their) and their analysis enables accurately clustering the recorded samples based on evolutionary feature similarity between the recorded samples and the seed samples. The number of reference samples is fixed and does not change, this being another system parameter (S).

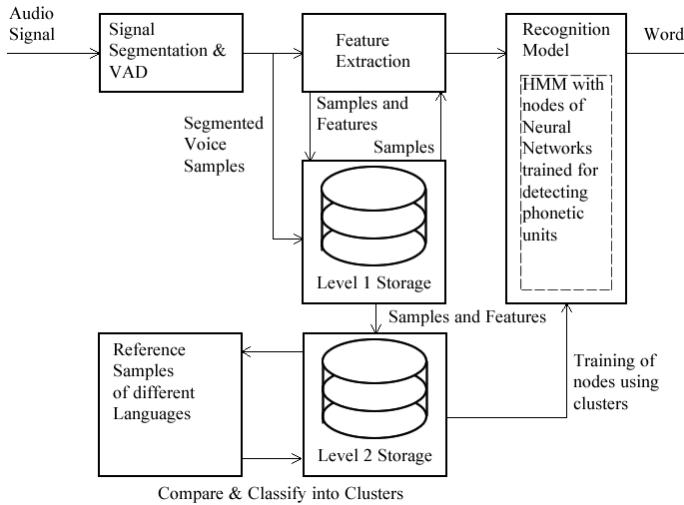


Fig. 2. Detailed view of the proposed structure

For feature extraction we will use computationally fast features currently employed in state of the art systems (wavelet transform, mel-frequency cepstrum coefficients) as well as a slower process that analyzes the distinctive features (consonant, sonorant, syllabic, laryngeal, manner and place features of phonetic units) of each sample beginning with the seed samples and then with the recorded samples in order to have a wide coverage of features for any given speech sample. To allow for such complexity we will have to use one neural network for each phonetic unit of the language. Clustering of recorded samples needed to train each detector will be made using a similarity determination algorithm based on a genetic algorithm that will compare the features of each recorded sound to those of each reference sample and classify the sample as belonging to the cluster of the most similar phonetic unit.

The fast feature extraction methods are used to obtain features from live speech which are then sent to the detectors that have also been trained with these types of features. The slow features are used for the clustering process, because they can't be a viable option when recognizing live speech since they are too slow to compute and because the clustering process has to rely on as much features as possible. When the storage limit is reached the feature extraction, clustering and training processes begin while the system stops recording and focuses on these computationally demanding operations.

In parallel with this acquisition algorithm there will be a standard feature extraction and recognition algorithm that uses the HMM detector network, this thread begins to run in parallel with the acquisition algorithm thread once the maximum number of iterations has been reached and has higher priority over the later when voice activity is detected.

Table 1. Proposed acquisition algorithm

```

Initialize L1, L2, S, IMAX parameters,  $i = 0$ 
WHILE ( $i <= \text{IMAX}$ )
  IF (Voice Activity Detected)
    Record sample to Level 1 Storage
  ELSE
    WHILE ( $\text{Level 1 storage} <= \text{L1}$ )
      Extract features
      IF ( $\text{Level 1 storage} = \text{L1}$ )
        Extract features from remaining recorded samples
        Move recorded samples to L2
        Clean Level 1 storage and
        Exit L1 WHILE
    WHILE ( $\text{Level 2 Storage} <= \text{L2}$ )
      Evaluate fitness of each sample with EA
      Cluster samples using seed samples features and EA
      IF ( $\text{Level 2 storage} = \text{L2}$ )
        Evaluate fitness of remaining samples
        Delete bottom half of each cluster according to
        fitness
        Exit L2 WHILE
    FOR (each c cluster from Level 2 storage)
      FOR (each sample  $s$  from sample cluster  $k$ )
        Train Neural Detector  $K$  with fast features of sample  $s$ 

```

5 Discussion of Results and Critical View

Our results consist of the proposed algorithm, structure and reference speech samples for the Romanian and English language as well as part of the implementation of the system using the Java programming language and for the hardware part we have two AKG professional microphones and a professional USB audio interface from M-Audio.

The presented algorithm and structure combined provide an improved paradigm for acquiring speech recognition, in an autonomous way, and a means to create new speech databases for training recognition models. We consider this as the starting point of our research into cognitive speech recognition and language acquisition for mobile robots.

6 Conclusions and Further Work

In this paper we have described a new paradigm for an automatic speech recognition system that mimics the acquisition of speech recognition capabilities by employing a novel structure and algorithm. Our research is in its incipient stage and therefore there is significant amount of testing and evaluation that remains to be done with our system.

Testing will have to validate the efficient and coherent acquisition of recognition knowledge as well as validating the recognition performances and capabilities of our solution by acquiring the knowledge to recognize Romanian and English languages by beginning with seed samples for each language.

The described structure and algorithms might potentially be improved and an optimum version must be found within the limits of our described three level framework. One way we could achieve this would be to optimize our system using genetic algorithms for identifying the optimum algorithms, neural networks models and parameters for our system. Another research possibility would be to use associative neural networks and self-organizing maps for the first two levels of storage instead of databases, and also implementing different types of neural networks for the selection and clustering processes in order to obtain an entirely neural networked based recognition system. Again optimizing this completely neural based structure would be possible by using evolutionary methods.

After we succeed in finding the best possible structure, algorithms, parameters and optimum settings for them we will proceed to designing and implementing a system that is capable of language acquisition while relying on our proposed system for continuous speech recognition.

Acknowledgement

This paper is supported by the Sectoral Operational Programme Human Resources Development (SOP HRD), financed from the European Social Fund and by the Romanian Government under the contract number POSDRU/88/1.5/S/59321.

References

1. Toscano, J.C., McMurray, B.: Cue Integration With Categories: Weighting Acoustic Cues in Speech Using Unsupervised Learning and Distributional Statistics. *Cognitive Science* 34, 434–464 (2010)
2. Dixon, P.R., Oonishi, T., Furui, S.: Harnessing graphics processors for the fast computation of acoustic likelihoods in speech recognition. *Computer Speech and Language* 243, 510–526 (2010)
3. Gales, M.J.F., Flego, F.: Discriminative classifiers with adaptive kernels for noise robust speech recognition. *Computer Speech & Language* 24, 648–662 (2010)
4. Jansen, A., Niyogi, P.: Point process models for event-based speech recognition. *Speech Communication* 51, 1155–1168 (2009)
5. Chater, N., Christiansen, M.H.: Language Acquisition meets Language Evolution. *Cognitive Science* 34, 1131–1157 (2010)
6. Hsu, A.S., Chater, N.: The Logical Problem of Language Acquisition. *Cognitive Science* 34, 971–1016 (2010)
7. Seitz, A.R., Protopapas, A., Tsushima, Y., Vlahou, E.L., Gori, S., Grossberg, S., Watanabe, T.: Unattended exposure to components of speech sounds yields same benefits as explicit auditory training. *Cognition* 115, 435–443 (2010)
8. Van der Velde, F., de Kamps, M.: Learning of control in a neural architecture of grounded language processing. *Cognitive Systems Research* 11, 93–107 (2010)
9. Lightfoot, D.: Language Acquisition and Language Change. *Wiley Interdisciplinary Reviews: Cognitive Science* 1, 677–684 (2010)
10. Casserly, E.D., Pisoni, D.B.: Speech Perception and Production. *Wiley Interdisciplinary Reviews: Cognitive Science* 1, 629–647 (2010)

11. Chen, S.-H., Guido, R.C., Truong, T.-K., Chang, Y.: Improved voice activity detection algorithm using wavelet and support vector machine. *Computer Speech and Language* 24, 531–543 (2010)
12. Shin, J.W., Joon-Hyuk Chang, J.-H., Kim, N.S.: Voice activity detection based on statistical models and machine learning approaches. *Computer Speech and Language* 24, 515–530 (2010)
13. Mohadese Eshaghi, M.R., Mollaei, K.: Karami Mollaei, Voice activity detection based on using wavelet packet. *Digital Signal Processing* 20, 1102–1115 (2010)
14. Jiang, H.: Discriminative training of HMMs for automatic speech recognition: A survey. *Computer Speech and Language* 24, 589–608 (2010)
15. Dede, G., Sazli, M.H.: Speech recognition with artificial neural networks. *Digital Signal Processing* 20, 763–768 (2010)
16. Majewski, M., Zurada, J.M.: Sentence recognition using artificial neural networks. *Knowledge-Based Systems* 21, 629–635 (2010)
17. Yu, D., Varadarajan, B., Deng, L., Acero, A.: Active learning and semi-supervised learning for speech recognition: A unified framework using the global entropy reduction maximization criterion. *Computer Speech and Language* 24, 433–444 (2010)
18. Wu, W.-L., Lu, R.-Z., Duan, J.-Y., Liu, H., Gao, F., Chen, Y.-Q.: Spoken language understanding using weakly supervised learning. *Computer Speech and Language* 24, 358–382 (2010)
19. Siniscalchi, S.M., Lee, C.-H.: A study on integrating acoustic-phonetic information into lattice rescoring for automatic speech recognition. *Speech Communication*
20. Nair, N.U., Sreenivas, T.V.: Joint evaluation of multiple speech patterns for speech recognition and training. *Computer Speech and Language* 24, 307–340 (2010)
21. Chien, J.-T., Chueh, C.-H.: Joint acoustic and language modeling for speech recognition. *Speech Communication* 52, 223–235 (2010)
22. Srinivasan, S., Wang, D.: Robust speech recognition by integrating speech separation and hypothesis testing. *Speech Communication* 52, 72–81 (2010)

HMM-Based Abnormal Behaviour Detection Using Heterogeneous Sensor Network

Hadi Aliakbarpour¹, Kamrad Khoshhal¹, João Quintas¹,
Kamel Mekhnacha², Julien Ros², Maria Andersson³, and Jorge Dias¹

¹ ISR, University of Coimbra, Portugal

{hadi, kamrad, jquintas, jorge}@isr.uc.pt

²Probayes, France

{kamel.mekhnacha, julien.ros}@probayes.com

³FOI, Linköping, Sweden

maria.andersson@foi.se

Abstract. This paper proposes a HMM-based approach for detecting abnormal situations in some simulated ATM (Automated Teller Machine) scenarios, by using a network of heterogeneous sensors. The applied sensor network comprises of cameras and microphone arrays. The idea is to use such a sensor network in order to detect the normality or abnormality of the scenes in terms of whether a robbery is happening or not. The normal or abnormal event detection is performed in two stages. Firstly, a set of low-level-features (LLFs) is obtained by applying three different classifiers (what are called here as low-level classifiers) in parallel on the input data. The low-level classifiers are namely Laban Movement Analysis (LMA), crowd and audio analysis. Then the obtained LLFs are fed to a concurrent Hidden Markov Model in order to classify the state of the system (what is called here as high-level classification). The attained experimental results validate the applicability and effectiveness of the using heterogeneous sensor network to detect abnormal events in the security applications.

Keywords: Heterogeneous sensor network, LLF (Low level Feature), HBA (Human Behaviour Analysis), HMM (Hidden Markov Model), LMA (Laban Movement Analysis), Crowd analysis, ATM (Automated Teller Machine) security.

1 Introduction

Recently, the demand for using automatic surveillance systems has been increasing. Many research areas, such as computer vision, signal processing, voice analysis and sensor fusion and pattern recognition are involved in this type of applications. Work on detection and tracking algorithms for dense crowds can be found in the literature. In [1] a method is suggested for simultaneously tracking all people in a dense crowd using a set of cameras with overlapping fields of view. In [2] a real-time system for detection of moving crowds is presented. HMM has been used in various applications for behavior recognition, e.g. [3] for facial action recognition and [4] for crowd behavior analysis. Crowd analysis can be used to get an understanding of the crowd as a whole, without any detailed information on individuals in the crowd. Crowd activity

provides information which can be used to detect if there are people running or fighting [4]. A deep contribution in the field of human-machine interaction (HMI), based on the concept of LMA (Laban Movement Analysis), is presented by Rett and Dias in [5]. In their work a Bayesian model is used for learning and classification. The LMA is presented as a concept to identify useful features of human movements to classify human gestures. Frequency-based extracted features are used in a LMA-based approach in our previous work for the sake of behaviour analysis [6]. Using audio signals for security purposes is proposed in [7]. Using distributed sensor network for the surveillance is investigated and proposed by Aliakbarpour et al. in [17,18].

A two-staged classification approach, to detect abnormal events in a security scenario, is introduced in this paper. Firstly, a set of low-level-features (LLFs) is obtained by concurrently applying three different classifiers (what are called here as low-level classifiers) on the input data. The low-level classifiers are namely LMA, crowd and audio analysis. Then the obtained LLFs are fed to a concurrent HMM in order to classify the state of the system (what is called here as high-level classification). A network of heterogeneous sensors such cameras and microphone arrays are used in a synergic way to observe the scene.

This paper is arranged as following: Our contribution to sustainability is presented in Sec. 2. Low-level classification is introduced in Sec. 3. A HMM-based classification (here is known as the high level classifier) is discussed in Sec. 4. Sec. 5 is dedicated to the experimental results and eventually the conclusion is presented in Sec. 6.

2 Contribution to Sustainability

In order to protect citizens, property and infrastructure, surveillance systems are increasingly being used. Commonly, these surveillance systems are installed in public spaces, covering large areas, where a great number of people populate the camera's fields of view. Consequently such systems consist often of a large amount of distributed sensors, typically CCTV cameras, monitored by operators in a control room. Since humans possess a limited capacity of driving its focus of attention, it is impossible to the system's operators pay attention to all what is happening in all monitors at a given time. Moreover, to recognizing abnormal or threatening events is a complex cognitive task requiring a focus that humans can uphold for only a short time. Therefore, there is the need for a persistent system capable of making a pro-active surveillance. In this paper we introduce a method to detect abnormal situation, in the ATM scenarios, using the observations of a heterogeneous sensor network comprising of cameras and microphone arrays.

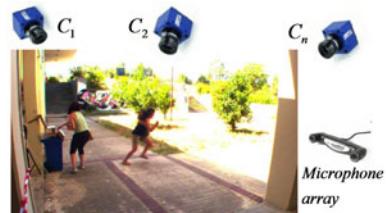


Fig. 1. A superimposed view of the ATM scenario when a robbery is happening

3 Low-Level Classification

As mentioned, the idea is to apply a two staged procedure in order to conclude whether the scene's state is normal or abnormal. Here the used data is from an exclusive

multimodal database, referred as PROMETHEUS database¹ [8]. This database is in support of the development and the evaluation of the algorithms which are intended to analyze and identify human actions and behaviors in the context of surveillance using multi-modal approaches. It comprises of various securities scenarios and among them we focused on the ATM ones. Many typical events such as walking, waiting, taking money and some atypical events such as robbery have happened in the ATM scenarios. The intention is to automatically identify and detect the state of the scene in terms of whether a normal or abnormal event is happening.

The scenes are observed by a network of heterogeneous sensors, composed of video cameras, thermal cameras and microphone arrays. Among the heterogeneous data in the database, we selected two modalities: image and sound. Three different low-level classification methods, namely LMA, crowd and audio analysis, have been applied on these data in order to obtain a set of LLFs. Table 1 summarizes the inputs and outputs for each method. As can be seen, two of these methods, the LMA and crowd analysis, have their inputs from applying some preliminary data fusion and tracking algorithms (which are supposed to be available) on the raw data and just the audio analysis one has a direct input from the raw sound signals. For the sake of having a low level classification on the sound signals, the method introduced in [7] has been used. Here we continue to introduce the LMA and crowd analysis methods.

Table 1. Inputs and outputs for the different methods in the low-level classification stage

Method	Input	Output
Crowd Analysis	Optical flow from image data	Crowd ratio
Laban Movement Analysis	3D positions of heads and feet (available from tracking algorithms)	$\text{Pr}(\text{walking}), \text{Pr}(\text{running}), \text{Pr}(\text{falling})$ and $\text{Pr}(\text{standing})$
Audio Analysis	Sound signals	LL-ratio

3.1 LMA-Based Human Movement Classification

LMA is a well-known method to describe and analyze human movements by several components; Body, Space, Shape, Effort and Relationship [9]. Each component describes human movements by different aspects. Among the different components of LMA, here we have selected Effort component to observe human movements in terms of how motion of human body parts are happening with respect to inner intention (our recent work in [6]). Effort has four sub-components and each of them has two states; Effort.time (sudden/sustained), Effort.space (direct/indirect), Effort.weight (light/strong) and Effort.flow (bounded/free). In this work, the Effort.time component for two body parts (head and feet) is estimated. Here the interesting movements to classify are standing, walking, running and falling-down. A Bayesian Network (BN), which is a well-known “tool” to deal with uncertainty, has been used for our LMA parameters estimations (to classify different human movements). The BN has three levels;

- The lowest level comprises frequency-based features which were achieved from Power Spectrum (PS) technique that applied on acceleration signals of body parts. First four coefficients were collected from PS signal of each body part acceleration

¹ www.prometheus-FP7.eu

signals. Each coefficient has four state possibilities (see [6]) which are defined by several thresholds.

- Middle-level includes LMA parameters, Effort.time component of the body parts.
- The highest level is just one node which has the number of interest human movement's states.

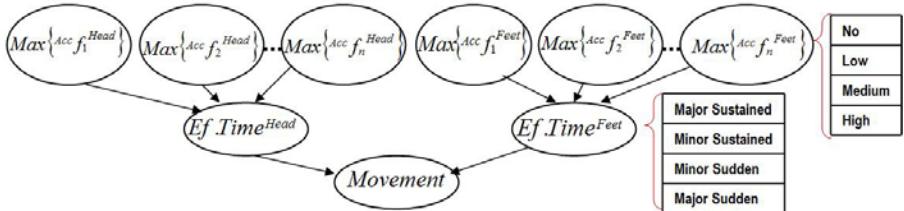


Fig. 2. LMA-based Bayesian net

3.2 Crowd Analysis

Crowd analysis is used to get an understanding of the crowd as a whole, without any detailed information on individuals in the crowd (see our recent work [4]). The aim is to get an approximate understanding of the activity level of the crowd as well as the size of the crowd. Crowd activity provides information which can be used to detect if there are people running or fighting. In general, normal crowd behavior often corresponds to calm movements, where people are standing or moving slowly through the scene, without making excessive gestures. Abnormal behaviour is instead likely to be accompanied by more rapid movements. The different levels of crowd activity are estimated by optical flow calculations, which estimate the relative motion between consecutive images. If a person is walking quickly, running, or moving his or her arms rapidly, the magnitude of optical flow will be larger compared to the case when a person is moving slowly or standing still. The different levels of crowd activity can be derived by using for example the following approaches:

1. Manually setting the different levels by observing the optical flow under known conditions, i.e. when the different types of events in the scene are known.
2. Applying a more automatic approach for obtaining the levels by using basic statistics (mean value and standard deviation) on relevant training data.

The crowd size provides information that can be used for getting warnings of forthcoming fights, attempted robbery and risk for riots. What is considered to be a large or small crowd will differ from case to case. For example, in a city area a large crowd may be 20-25 persons or more. At large sport events, large crowds are probably hundreds or thousands of persons. Fighting and robbery at the ATM means that at least two persons are present on a small area at the same time. The crowd size estimate is obtained by first performing background subtraction to obtain the foreground pixels. We assume prior knowledge of the approximate number of pixels per person, which depends on the distance between camera and crowd. The number of people is then obtained by dividing the total amount of foreground pixels by the number of assumed pixels per person. Also crowd growth rate and crowd density can be of interest to

understand the crowd behaviour. Is the crowd growing quickly and/or are the people standing close to each other? Crowd growth rate can be estimated by studying the change in crowd size for a certain reasonable time period. Crowd density can be estimated by relating the crowd size to a specific area in the image. By combining estimates of crowd activity and crowd size, uncertainty in the classification of behavior will be reduced. For example, increased activity (a running person) at an ATM *together* with the information that there were at least two persons present close to the ATM, will strengthen the view that there have been an attempted robbery.

4 Concurrent HMM-Based Classification

For behaviour recognition, we are interested in detecting the current behavior amongst N known behaviors (i.e. the behaviour library). For this purpose, we propose to use a concurrent HMM architecture.

Principle. A concurrent hidden Markov model is composed of several hidden Markov Models, each one describing one class (see Fig.3-left). To summarize, the concurrent HMM centralizes the on-line update of the behaviour belief and contains:

1. The set of HMMs representing basic behaviour library (one HMM per behaviour);
2. The transition between behaviors model that could be either defined by hand (by an expert), or learnt from annotated data.

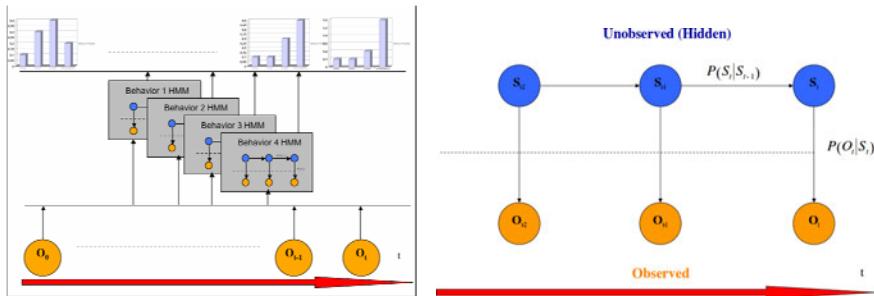


Fig. 3. Concurrent Hidden Markov Models (left) HMM model (right)

Hidden Markov Models are used to characterize an underlying Markov chain which generates a sequence of states. The term Hidden in the HMM name comes from the fact that the sequence of states is not directly observable. Instead the states generate an observable sequence. Thus, the output depends on the current state and on previous outputs. These tools are widely used in the field of sound processing [11], gene finding and alignment in DNA sequences [12]. They were introduced by Andre Markov in [13] and developed in [14].

HMM are widely employed in the field of computer vision to recognize gesture or human behaviour [15]. In these applications, the observation variables are features extracted for video data. The principle of an HMM is presented on Figure

Fig.3-(right) in which S represents the state variable and O represents the observation ones. In our application, each HMM describes a human behaviour and is learned using a training dataset composed of labeled observation sequences that are low-feature extracted from the different and thanks to three different approaches, namely Laban Movement Analysis (LMA), crowd and audio analysis.

Construction/Learning. Constructing a concurrent hidden Markov model consists in:

- Learning the set of HMM models representing the behaviour library (one HMM per behaviour) using an annotated data set.
- Defining the transition matrix between the behaviors. This transition model could be either defined by hand (by an expert), or learnt from an annotated data set.

Learning the behaviour transition model is straightforward and consists in computing simple statistics (histograms) of transitions using the annotated data set.

Learning the underlying HMM models (a HMM per behaviour) is more complex. It can be divided into two sub-problems:

1. Finding the optimal number of states N . The optimal number of internal states within the HMMs could be chosen by hand thanks to an expert. In this case no algorithm is needed and the learning of the HMM is reduced to the learning of its parameters. However, since an HMM is a Bayesian Network, a score that allows a compromise between fitting learning examples (D) and the ability of generalization (see the Occam Razor Principle) can be employed to find it automatically [15]. For example the classical Bayesian Information Criterion [16] that maximizes the likelihood of the data while penalizing large size model can be used:

$$BIC(n, D) = \log(\text{likelihood}(D, n)) - \frac{1}{2} \times n\text{params}(n) \times \log(|D|)$$

In this case the optimal number of states is given by: $n^* = \arg \max_n BIC(n, D)$.

2. Learning the parameters of the HMM given N (i.e., the transition matrix $P(S_t | S_{t-1})$, the observation distribution $P(O_t | S_t)$, and the initial state distribution $P(S_0)$). The idea is find the parameters that maximize the data likelihood. For this purpose the methods generally employed are the classical EM algorithm (aka Baum-Welch algorithm in the HMM context), or the Iterative Viterbi algorithm.

Recognition. As previously emphasized, the concurrent hidden Markov model is used to recognize on-line or off-line the current behaviors amongst N known behaviors. This is easily performed by finding the HMM M that maximizes $P(M | O_{t-n}, \dots, O_t)$ for the off-line case (or $P(M | O_t)$ for the on-line case).

5 Experiments

As mentioned, the ATM scenarios of PROMETHUS [8] multimodal database are used in this paper. There are four selected ATM scenes with different durations. The interesting event which we call as abnormal state in this kind of scenario is robbery.

Based on the proposed approach the process of event detection has two levels for performing the classification. Firstly, some low-level features are obtained by using three different approaches (see Fig.3). Then these low-level features are fed to a HMM as a high level classifier in order to estimate the scene's state. Apart of these LLF inputs for the HMM, another parameter which is named as “*environment parameter*” is also consider for the HMM. The *environment parameter* is defined as the relative positions of people to the ATM. In the context of the ATM scenario in PROMETHEUS dataset, we defined the robbery state as when the robber waits in ATM's area, approaches a person who is taking money from the machine, steals the money and then rapidly escapes. In the database, there are 139 samples corresponding to the normal situations and 8 samples corresponding to the robbery (abnormal) situations. Each sample has a 10 seconds long. Among these samples, 61 samples of normal data and 4 samples of abnormal ones have been randomly selected for HMM learning process and the others (78 samples of normal and 4 samples of abnormal) for HMM classification process.

Here we have implemented different experiments on the data in order to demonstrate the applicability and effectiveness of using heterogeneous data fusion in the proposed manner. Table 2 depicts the result of the HMM-based classifier when the output of each low-level classifier is individually applied to the high-level (final) classifier. Then the experiment is performed when a pair of the low-level-classifier outputs is used for the classification (three possible pair combinations, see Table 3). Eventually all of the low-level classifier's outputs have been fed to the high-level classifier. Table 4 depicts the result of this last case, in which there are the best percentages of the true even detections. It validates the effectiveness of the proposed method for the sake of surveillance applications.

Table 2. High-level classification result: when just one of the three low-level features has been used

Method	LMA			Crowd			Sound(LL ratio)		
	Normal	Robbery	%	Normal	Robbery	%	Normal	Robbery	%
Normal	72	6	92	64	14	82	76	2	97
Robbery	0	4	100	1	3	75	1	3	75

Table 3. High-level classification result: Three possible combinations of using a pair of low-level features

Methods	LMA + Crowd			LMA + Sound (LL ratio)			Sound (LL ratio) + Crowd		
	Normal	Robbery	%	Normal	Robbery	%	Normal	Robbery	%
Normal	72	6	92	77	1	98	77	1	98
Robbery	0	4	100	0	4	100	1	3	75

Table 4. High-level classification result: Fusion of all low-level features

Method	LMA + Sound (LL ratio) + Crowd		
	Normal	Robbery	%
Normal	77	1	98
Robbery	0	4	100

6 Conclusion

Abnormal human behavior detection by using a network of heterogeneous sensors, in some ATM scenarios, has been proposed in this paper. A two-staged classification, divided as low-level-classification and high-level-classification, is used in order to detect the abnormality of the current state of the scene. Here, the interesting abnormal event is defined as happening a robbery near the ATM. The LMA, crowd analysis and audio analysis are the methods which are used in the low-classification stage. For the sake of high-level classification, a concurrent HMM is applied. The attained experimental results validate both the applicability and efficiency of the proposed method for the sake of surveillance applications.

Acknowledgment. *Hadi Ali Akbarpour* is supported by the **FCT** (Portuguese Fundation for Science and Technology). This work is supported by the European Union within the FP7 Project **PROMETHEUS**, www.prometheus-FP7.eu. The authors would like to thank our partners from University of Patras for providing sound data.

References

- [1] Eshel, R., Moses, Y.: Homography Based Multiple Camera Detection and Tracking of People in a Dense Crowd. In: IEEE Conference on CVPR 2008 (2008)
- [2] Bird, N., Atev, S., Caramelli, N., Martin, R., Masoud, O., Papanikolopoulos, N.: Real time, online detection of abandoned objects in public areas. In: Robotics and Automation, ICRA 2006, May15-19, pp. 3775–3780. IEEE, Los Alamitos (2006)
- [3] Shang, L., Chan, K.-P.: Nonparametric discriminant HMM and application to facial expression recognition. In: CVPR 2009, June 20-25, pp. 2090–2096. IEEE, Los Alamitos (2009)
- [4] Drews, P., Quintas, J., Dias, J., Andersson, M., Nygards, J., Rydell, J.: Crowd behavior analysis under cameras network fusion using probabilistic methods. In: The 13th International Conference on Information Fusion, EICC Edinburgh, UK, July 26-29 (2010)
- [5] Rett, J., Dias, J., Ahuactzin, J.-M.: Laban Movement Analysis using a Bayesian model and perspective projections. Brain, Vision and AI (2008)
- [6] Khoshhal, K., Aliakbarpour, H., Quintas, J., Drews, P., Dias, J.: Probabilistic LMA-based classification of human behaviour understanding using power spectrum technique. In: 13th Int. Conf. on Information Fusion 2010, EICC Edinburgh, UK, July 10 (2010)
- [7] Ntalampiras, S., Potamitis, I., Fakotakis, N.: An Adaptive Framework for Acoustic Monitoring of Potential Hazards., EURASIP. Journal on Audio, Speech, and Music Processing Volume (2009) doi:10.1155/2009/594103
- [8] Ntalampiras, S., Ganchev, T., Potamitis, I., Fakotakis, N.: Heterogeneous Sensor Database in Support of Human Behaviour Analysis in Unrestricted Environments: The Audio Part The 7th int. conf. on Language Resources and Evaluation, LREC (2010)
- [9] Zhao, L., Badler, N.I.: Acquiring and validating motion qualities from live limb gestures. Graphical Models, 1–16 (2005)
- [10] Shi, G., Zou, Y., Jin, Y., Cui, X., Li, W.J.: Towards HMM based human motion recognition using mems inertial sensors. In: Proc. IEEE Int. Conf. Robotics & Biomimetics (2009)

- [11] Rabiner, L.R.: A tutorial on Hidden Markov Models and selected applications in speech recognition. pp. 267–296 (1990)
- [12] Pachter, L., Alexandersson, M., Cawley, S.: Applications of generalized pair Hidden Markov Models to alignment and gene finding problems. In: Proceedings of the 5th Annual Int. Conf. on Computational Biology, RECOMB 2001, New York, USA, pp. 241–248 (2001)
- [13] Markov, A.: An example of statistical investigation of the text eugene onegin concerning the connection of samples in chains. In: Lecture at the physical-mathematical faculty, Royal Academy of Sciences, St. Petersburg
- [14] Baum, L.E., Petrie, T., Soules, G., Weiss, N.: A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics* 41(1), 164–171 (1970)
- [15] Biem, A.: A model selection criterion for classification: Application to HMM topology optimization. In: Proceedings of the Seventh International Conference on Document Analysis and Recognition, ICDAR 2003, Washington, DC, USA, p. 104. IEEE, Los Alamitos (2003)
- [16] Schwarz, G.: Estimating the dimension of a model. *The Annals of Statistics* 6(2), 461–464 (1978)
- [17] Aliakbarpour, H., Ferreira, J.F., Khoshhal, K., Dias, J.: A Novel Framework for Data Registration and Data Fusion in Presence of Multi-modal Sensors. In: Camarinha-Matos, L.M., Pereira, P., Ribeiro, L. (eds.) DoCEIS 2010. IFIP Advances in Information and Communication Technology, vol. 314, pp. 308–315. Springer, Heidelberg (2010)
- [18] Aliakbarpour, H., Dias, J.: Human Silhouette Volume Reconstruction Using a Gravity-based Virtual Camera Network. In: The Proceedings of the 13th International Conference on Information Fusion 2010, EICC Edinburgh, UK, July 26-29 (2010)

Displacement Measurements with ARPS in T-Beams Load Tests

Graça Almeida^{1,2}, Fernando Melicio², Carlos Chastre¹, and José Fonseca¹

¹ Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa

² Instituto Superior de Engenharia de Lisboa,
Lisboa, Portugal

mdg.almeida@fct.unl.pt, fmelicio@deea.isel.ipl.pt,
chastre@fct.unl.pt, jmf@uninova.pt

Abstract. The measurement of deformations, displacements, strain fields and surface defects in many material tests in Civil Engineering is a very important issue. However, these measurements require complex and expensive equipment and the calibration process is difficult and time consuming. Image processing could be a major improvement, because a simple camera makes the data acquisition and the analysis of the entire area of the material under study without requiring any other equipment like in the traditional method. Digital image correlation (DIC) is a method that examines consecutive images, taken during the deformation period, and detects the movements based on a mathematical correlation algorithm. In this paper, block-matching algorithms are used in order to compare the results from image processing and the data obtained with linear voltage displacement transducer (LVDT) sensors during laboratorial load tests of T-beams.

Keywords: Digital Image Correlation, Block Motion Estimation, image processing.

1 Introduction

Since the 80's, several works about digital image correlation techniques are under development in order to obtain an accurate knowledge of the displacement field [1-3]. The general idea behind digital image processing techniques is to calculate the displacement field or the strain field without contact and using a simple low cost camera. The research aims to reduce the computation time and to increase the accuracy of the system. When conventional methodology is used the number of the measured points takes a huge importance because they increase the hardware, the time to get the setup ready and the costs. Using image analysis techniques the density of the measured points can be very high. As an example, a trivial image of 1024 by 1024 pixels can be used to obtain a continuous information field with more than 4000 analysis points.

In our previous work[4] the three-step search (TSS) [5] algorithm was studied, especially the simple and efficient (SES) algorithm. A partnership between researchers from the Civil Engineering Department and the Electrical Engineering Department, of the Universidade Nova de Lisboa (UNL) made possible to explore the traditional methodology used in Civil Engineering measurements, which requires a complex and expensive sensorial setup and a very complex calibration process. In this work we compare the results of the traditional approach with the results obtained by digital image processing techniques.

In this paper the purpose is to use the adaptive rood pattern search (ARPS) and compare it with the data obtained with physical sensors [6]. In this paper only two concrete beams are compared. Although concrete has by itself a good texture a speckle pattern was applied to the T-beams surfaces in order to get an easier image processing (in this paper a random dot pattern was applied). The work developed so far shows that resolution versus specimen dimensions, focal length, distance between camera and specimen, distortion and the speckle pattern are some of the factors that influence the most the measurements error. In order to choose the most promising correlation technique several tests were conducted, such as the comparison of different block matching algorithms and the use of the edge detection techniques. As shown in the previous work, [4] the TSS algorithm shows to be more adequate than edge detection techniques. However, some research questions such as what kind of correlation technique is more efficient or how to increase the accuracy of the system without increasing computation time are still open. Another open issue is what should be the relationship between the entire image and the number of target points. The comparison between the image processing results and the information obtained by the LVDT sensors will be used to establish this relationship.

In [7] a non-touching strain measurement method that covers a pre-defined area is presented. In this work several tests were carried out on RC beams with a span of 4.5 m and the results were compared to the traditional electrical strain gauges. A speckle pattern correlation was used and a photo was taken every 30kN until rupture. In this work a camera film was used and the film was scanned with a high quality scanner. The picture was divided into 128x128 sub-pictures and then subjected to a threshold. The centre of gravity of the resulting black & white picture was then calculated. With this methodology and the speckle correlation it was possible to find the same sub-picture in the second loading condition and therefore estimate the deformation. Despite the good results obtained by this methodology it is significantly dependent on the threshold value and therefore varying with the illumination conditions, speckle pattern and image acquisition parameters. In our work we look for a more stable and conditions independent solution for this problem.

2 Contributions to Sustainability

Image processing analysis is used in many situations for earth observations such as resource managements, soils mapping, water resources, etc. The measurements made by image processing simplify the maintenance of local materials, the waste or loss of sensors and human resources are best applied.

With measurements without contact, fewer materials are used which implies reduction of costs and energy. In our tests the same system could analyze several materials with no duplication of the sensors. So it is possible to contribute to a better world with reduction of plastics, iron, and energy.

3 ARPS - Block Motion Algorithm

The basic idea of block motion estimation is to divide the current image into a matrix of blocks and then compare these blocks with the previous image in order to calculate

the displacement vectors. The current block is searched in the previous image in a delimited search area, p pixels around the current block. In our tests a block takes a square side of 16 and the search parameter p is 7.

Block-based motion estimation assumes that objects move in a translational movement. In (Fig. 1) it is possible to see an example of two images that could be compared. At left is the non deformed image and at right the deformed image. It is also possible to see the grid of 16 x 16 pixels and the random pattern that was applied. The methodology for estimation of displacement uses the intensity of the pixel and with the unique pattern in each sub-region makes it possible to find the displacement.

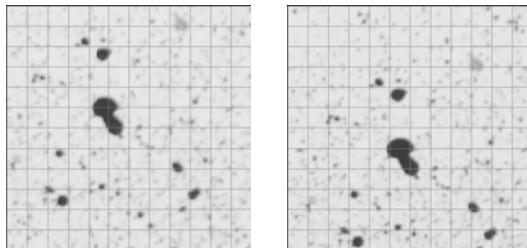


Fig. 1. At left is a no deformed image and at right is the deformed image; images are sub-divide in sub region of 16x16 pixels

The match block algorithm chosen is based on the least cost. The cost function normally used is the mean square error (MSE) or the mean absolute difference (MAD). In this work the MAD function was used in order to save computation time.

In most cases adjacent blocks have similar motions. The block on the immediate left, above, above-left and above-right of the current block are the most important to calculate the predicted the motion vector (MV) [8]. Four types of region of support could be used (Fig. 2). In this paper the type D was chosen because it requires less memory.

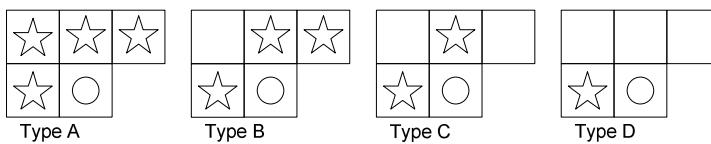


Fig. 2. Regions of support: the blocks marked by “O” are on each case the current block and the blocks marked with a star are used for predicted the MV

For the initial search, the ARPS algorithm evaluates the four endpoints in a symmetrical rood pattern plus the predicted MV. The four arms of the rood pattern are of equal length (Fig. 3).

The size of the rood pattern is equal to the length of the predicted motion vector (i.e. the motion vector of the immediate left of the current block). In each search points is necessary to compute the MAD function. The size of the rood pattern, Γ , is calculated in (1),

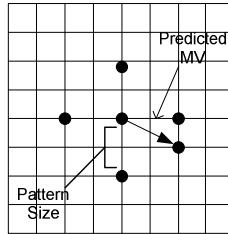


Fig. 3. Symmetrical shape of Adaptive Rood Pattern with four search points locating at the four vertices

$$\begin{aligned}\Gamma &= \text{round} \left| \vec{MV}_{predicted} \right| \\ &= \text{round} \left[\sqrt{MV_{predicted(x)}^2 + MV_{predicted(y)}^2} \right]\end{aligned}\quad (1)$$

The square and the root square operations drawn in (1) require a lot of computation time. Therefore instead of (1) it is possible to use a simplification that only requires the highest magnitude of the two components of the predicted MV (2).

$$\Gamma = \text{Max} \left\{ \left| MV_{predicted(x)} \right|, \left| MV_{predicted(y)} \right| \right\} \quad (2)$$

When it is not possible to apply the type D of the ROS, the value 2 is chosen for the size of the arm length (i.e. $\Gamma=2$). The minimal matching error (MME) point found in the current step will be re-positioned as the new search center of the next search iteration until the MME point is found as the center of the fixed pattern.

4 Results

Several load tests have been carried out in order to compare the results obtained by the image processing techniques with the information acquired by a classical measurement system. These tests [6] uses two reference T-beams and three T-beams strengthened with different FRP techniques. In this paper only the results from TSC1 and HB2 are used. The image acquisition conditions for these tests are shown on Table 1. Each image is divided in sub regions of 16×16 pixels ($\sim 4.4 \text{ mm} \times 4.4 \text{ mm}$).

The TSC1 beam was used as reference without any additional flexural reinforcement and the HB2 beam was reinforced with 3 bonded GFRP sheet layers. The T-beams had a 3m span by 0.3m heights and were tested until rupture in a 4-point

Table 1. T-Beam image acquisition condition: resolution and the number of photos

T beam designation	Resolution	Number of Photos
TSC1	36 pixel/cm	52
HB2	35 pixel/cm	27

bending test system. All the tested beams followed a monotonic loading history. The deflection control was granted by seven standard 100mm LVDT, displayed along the longitudinal direction of the beam. At the mid-span, a wire-controlled transducer (500mm) was used together with the LVDT (100mm) in order to obtain the results of larger deformation, if needed during the post-collapse of the test. The data from the LVDT at the mid-span was used for the comparison with the data obtained with the image system analysis.

The image acquisition for TSC1, and HB2 was done with a digital Cannon EOS 400D camera with a resolution of 3888x2592 and two spots of 500W each guaranteed artificial light. The artificial lightning was used in order to maintain a constant light environment. All the images were captured on RAW format and then convert to TIFF format for image processing with MATLAB.

Before the data acquisition it was necessary to make the preparation for the image acquisition system (Fig. 4) and the sensor data system acquisition.

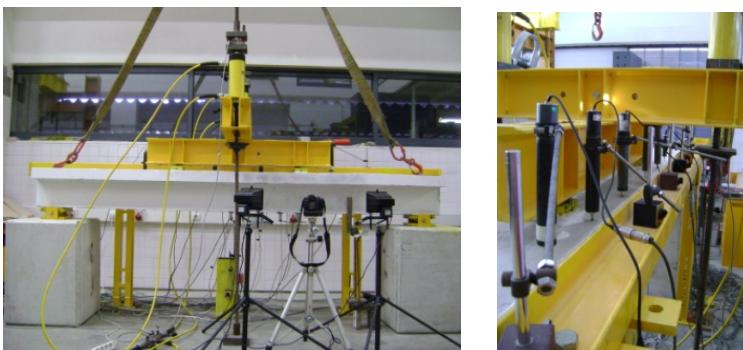


Fig. 4. The preparation of the image acquisition system (left) and data sensor system (right)

The T-beam was initially prepared with an underlying painting of white mate ink and then with a superimposed random speckle pattern manually painted using a large brush with mate black ink. All the digital measurements are done at a distance without any particular or long calibration, having a low cost support, and they are easy to implement in the T-beam test setup.

The photos were taken with intervals of 30 seconds. In first image of the test, the area of interest is over and at the end is down (Fig. 5). In some load tests the displacement could be large.

The digital image processing system compares two adjacent images in order to evaluate the displacement. These results are compared with the results acquired with one of the real sensors used and this evaluation can be seen in a graph of the displacement versus time (d vs. t).

In a non-contact strain measurement it is possible to analyze a small area of interest and a large one with the same hardware conditions. Moreover, this technique can be compared with non-linear software based on the finite element analysis (FEA) in order to corroborate the strain fields in the T-beam structure.



Fig. 5. Images of the test: the first image (left) and the last image (right)

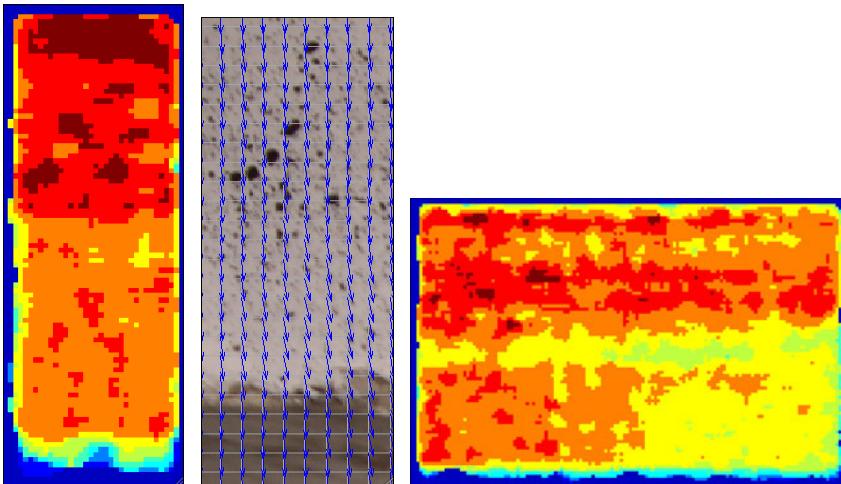


Fig. 6. Representation of displacement map (left) where the dark regions indicated larger displacements, the vector diagram (centre) shows the flow of deformation and with the same system is also possible to compute the complete displacement map (right)

The strain map and the grid vector of the movement vector displacement were done for the small part associated with the LVDT at the mid-span (Fig. 6).

The displacement vector diagram shows the compressive vertical load applied to the specimen. In order to know the implication of the entire T-beam we have also done the strain map of the entire T-beam.

The ARPS algorithm was used in the region of the image associated with the position of the LVDT, for each experiment (TSC1 and HB2), it is possible to visualize the evaluation of the displacement versus time (d vs. t) (Fig. 7).

The abrupt jump indicates that a major crack is found and in this situation the image processing has some problems. The error, before crack situation, was 0.4 mm for TSC1 and 4.3 mm for HB2. The error can be reduced if we reduced the interval between photos.

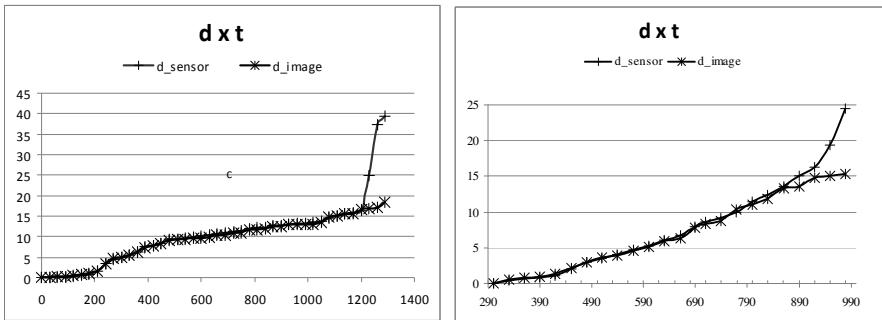


Fig. 7. Evaluation of displacement vs. time with results from image processing (**) point) and LVDT sensor (++ point): TSC1 at left and HB2 at right

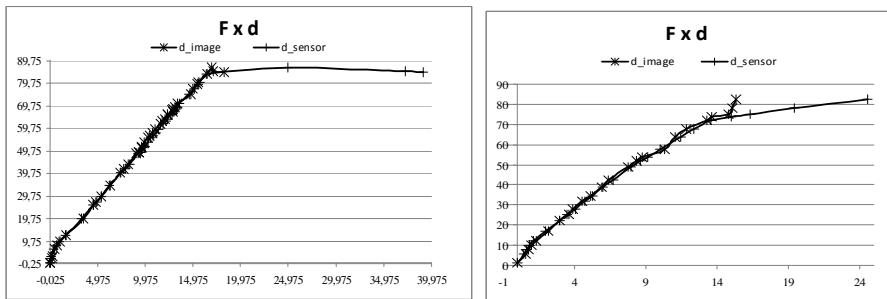


Fig. 8. Evaluation of the force vs. displacement: TSC1 at left and HB2 at right. Values from sensor are marked with (++) and image processing with (**).

The evaluation of force versus displacement (F vs. t) is also shown (Fig. 8) and it is also possible to visualize that our system is close to the real sensors.

5 Conclusions

The comparison between the measurements from LVDTs sensors and the measurements obtained from image processing techniques revealed to be very similar. Thus, the image processing technique used seems to be a very promising technique for measurement displacements with a much lower investment and much faster and easier setup.

The results obtained in this study shows that it is possible to continue and explore this subject and test the whole area of a T-beam in order to compare with the remain LVDTs sensors. Moreover, a strain field of the T-beam in study can be reproduced and compared to other techniques of modeling structures such as FEA, in order to calibrate and validate the image processor technique. The research will eventually provide software that enables a real time monitoring of the RC structure during the experimental test. In spite of using median and high pass filter amongst others pre-processing image technique, the results does not show a significant improvement.

It is important in future work to analyze the influence of the regular pattern and also increase the resolution of the image acquisition in order to obtain more precise results.

References

1. Chu, T.C., Ranson, W.F., Sutton, M.A., Peters, W.H.: Applications of Digital Image-Correlation Techniques to Experimental Mechanics. *Experimental Mechanic* 25(3), 232–244 (1985)
2. Sutton, M.A., Wolters, W.J., Peters, W.H., Ranson, W.H., McNeill, S.R.: Determination of displacements using an improved digital correlation method. *Image and Vision Computing*, 133–139 (1983)
3. Peters, W.H., Ranson, W.: Digital Imaging Techniques In Experimental Stress Analysis. *Optical Engineering* 21(3), 5 (1982)
4. Almeida, G., Biscaia, H., Melicio, F., Chastre, C., Fonseca, J.: Displacement Estimation of a RC Beam Test based on TSS algorithm. In: 5th Iberian Conference on Information Systems and Technologies (CISTI 2010), Santiago de Compostela (2010)
5. Barjatya, A.: Block Matching Algorithms for motion Estimation, DIP 6620 Spring, Final Project Paper (2004)
6. Carvalho, T., Paula, R., Biscaia, C.C.: Flexural Behaviour Of Rc T-Beams Strengthened With Different Frp Materials. In: 3rd Fib International Congress, Washington, D.C (2010)
7. Carolin, A., Olofsson, T., Taljsten, B.: Photographic strain monitoring for civil engineering. In: FRP Composites in Civil Engineering - CICE 2004, Seracino (2004)
8. Yao Nie, K.-K.M.: Adaptive Rood Pattern Search for Fast Block-Matching Motion Estimation. *IEEE Transactions On Image Processing* 11(12), 8 (2002)

Wireless Monitoring and Remote Control of PV Systems Based on the ZigBee Protocol

V. Katsioulis¹, E. Karapidakis², M. Hadjinicolaou¹, and A. Tsikalakis²

¹ School of Engineering and Design, Brunel University, UB8 3PH, Uxbridge, UK

² Renewable Energy Engineering Lab, TEIC, Romanou 3 str, 73133 Chania, Greece

Abstract. Systems that convert the sunlight into electrical energy like photovoltaics (PV) have been becoming widespread worldwide. The prospect of using the promising technology of wireless sensor networks (WSN) in the field of PV plant supervising and monitoring is studied here. The knowledge of the status and good working condition of each PV module separately as well as of any PV system component will lead in a more efficient way for power management. The nature of the wireless sensor networks (WSN) offers several advantages on monitoring and controlling applications over other traditional technologies including self-healing, self-organization and flexibility. The versatility, ease of use and reliability of a mesh network topology offered by the ZigBee technology that is based on the IEEE 802.15.4 standard, is used here to offer its maximum advantages on a system that is capable for real time measurements and event alerts.

Keywords: Photovoltaic panels, PV monitoring & control, wireless sensors networks and ZigBee.

1 Introduction

Many monitoring (data-acquisition) systems have been proposed and developed. The main task of these systems is to monitor and collect data concerning the performance of the PV plant in real time. In [1] a system for remote monitoring and control of complex stand-alone photovoltaic plants is proposed. It is based on the NI Field-Point architecture, an FP-2000 controller and an FP-I-100 acquisition system. Current, voltage and temperature data are transmitted using a GSM modem. A different approach has been proposed in [2], [3] where a computer-based data-acquisition system monitors both electrical and meteorological data. Remote clients can reach data using the TCP/IP protocol. Another proposal of PV monitoring is described in [4].

The above proposals give only a general performance monitoring image. They do not provide information about the performance and state of each individual PV module. In some cases, a specific PV module in a large scale PV plant may produce no or lower energy levels than expected. In this case, in a monitoring system architecture similar with those described above, the system will be able to sense and monitor the lower current or voltage reading but it has no ability to locate the source of the problem. This is mainly because current and voltage sensors are connected at

the output of a PV array that might be part of several PV modules. In this case it is obvious that a maintenance operator is needed to manually locate the defective part.

Wireless sensor networks (WSN) are a very promising technology in the field of PV monitoring. A wireless sensor network is a system which comprises radio frequency (RF) transceivers, sensors, micro-controllers and power sources. A sensor measures physical parameters like temperature, light, pressure, voltage and current. Wireless sensor networks with self-organizing, self-configuring, self-diagnosing and self-healing capabilities have been developed to solve problems or to enable applications that traditional technologies could not address [5]. A major advantage of this kind of networks is low cost, small size and low power consumption. In the field of WSN, the ZigBee technology had met a wide acceptance because of its capability to operate in a large number of applications.

In this paper, the main objective is to study the functionality of a ZigBee based monitoring and supervising system. The investigated PV system is located on the roof of the Technological and Educational Institute of Crete (Chania/Greece). It consists of six (6) polycrystalline-Si technology PV modules and rated power 100 Wpeak each connected in parallel. The remainder of this paper deals with a brief description of the ZigBee standard, a presentation of the developed hardware and the current case study with representative experimental measurements.

2 Contribution to Sustainability

Studies on photovoltaic phenomena have drawn the attention to the problem of PV behavior under varying environmental conditions [6]. The major problem is the strong dependence of a PV system response on many extrinsic factors such as temperature, insolation, cloudiness and pollution [7]. Another problem to be solved is to find an efficient and also cost-effective monitoring and supervising method of a PV plant even for small PV systems. Regular performance checks on the functioning of PV systems are necessary for a reliable use and successful integration of a PV. Monitoring for large PV systems is performed by specially designed hardware and software that might be expensive and is mainly operated by specially trained personnel. For small PV systems, up to 10kW, these checks are not performed due to monitoring system costs [8]. Therefore such small systems are often not checked on a regular basis. This situation can lead to partial energy losses that usual originate from partial system faults or decreasing performance that can be unnoticed for a long time. Thus, in order to achieve the maximum energy out of a PV plant, a low-cost, easy to use monitoring system is needed.

ZigBee technology is a low data rate, low power consumption, low cost, wireless networking protocol targeted towards wireless connections between electronic devices in automation and remote controlling applications [9, 10]. ZigBee can be implemented in mesh networks larger than it is possible with Bluetooth. ZigBee compliant wireless devices are expected to transmit 10-50 meters, depending on the RF environment and the power output consumption required for a given application [11-14].

3 ZigBee and System Architecture

The ZigBee standard is built on top of the IEEE 802.15.4 standard. The IEEE 802.15.4 standard defines the physical and MAC (Medium Access Control) layers for low-rate wireless personal area networks [18, 19]. Other benefits of this standard are energy measurements over the operated channel, Link Quality Indication (LQI), Received Signal Strength Indication (RSSI) and clear channel assessment. Different network topologies are supported like tree, star, peer to peer and mesh. ZigBee is targeted mainly for battery-powered applications where low data rate, low cost, and long battery life are the main requirements. The main field of applications for ZigBee focuses on industrial control, home automation, energy monitoring, wireless logistic systems and many others, as this standard tends to meet wide acceptance from semiconductor companies.

In this study, PV monitoring system architecture consists of two (2) basic blocks; the PV area sector (PVAS) and the central station sector (CSS):

PV area sector: This is actually the PV plant area where basic monitoring parameters are monitored. An alternative title instead of PVAS could be the (ZigBee modules area) and this is because each PV module in the plant is equipped with a ZigBee module and a set of sensors. The PVAS is part of ZigBee end devices (ZED) and a ZigBee router node (ROUT) that serves all ZED's as a sink to the ZigBee coordinator (COO).

Central Station sector: It can be assumed that CSS is the control station of a PV system. In such a station, various components of a PV system like the inverter, the batteries and the battery charger could be found. In the CSS the ZigBee coordinator module, the host PC and the remote measurement board (REMB) are located.

Remote Measurement Board: For safety reasons, all high voltages and currents of the PV plant are monitored in a separate PC board that is connected to the coordinator board.

Furthermore, the monitoring system generally operates as follows:

1. The attached in each PV module ZigBee notes, sends data about their voltage, current and temperature volumes. Data from ZEDs are sent to the router which is also an attached mote to a PV module. The router has the same functionality as with any ZED but is additionally capable of PV array angle reading (when a sun tracking system is used) and has also a dust sensor. All collected data from ZEDs and the router itself are sent to the coordinator board.
2. The coordinator board collects data from ZEDs via the router and supplies the host PC. The host PC runs suitable Graphic User Interface (GUI) software where monitoring and data logging are implemented.

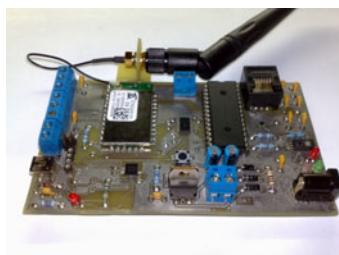
Finally, the coordinator board monitors the total current and voltage originating from the PV plant and inverter as well. Extra functionality such as inverter and battery temperature as well as relative humidity is available. In the following Table 1 a list of all the measured parameters by the developed PV monitoring system are presented.

Table 1. List of the monitored parameters

No	Parameter	No	Parameter
1	Modules output voltage	8	Inverter output current
2	Modules output current	9	Inverter temperature
3	Modules temperature	10	Battery voltage
4	Ambient temperature	11	Battery charging current
5	System output voltage	12	Battery temperature
6	System output current	13	Relative humidity
7	Inverter output voltage	14	Dust air concentration

3.1 Coordinator Board

As described above, the COO board is part of the CSS. The coordinator board is the central component of the monitoring system due to the fact that it is the coordinator of the ZigBee network. The coordinator board that depicted in Fig. 1 hosts the ZigBee module (ETRX2-PA), an Analog to Digital Converter (ADC) (ADC0808), a Universal Serial Bus (USB) to RS-232 interface/converter (CP2102), a battery backed up power supply, a relative humidity sensor (HIH-3041) and an RJ-45 for connection to the REMB.

**Fig. 1.** The coordinator board

A USB port is provided for connection of the COO board to the host PC. Relative humidity levels as well as the control station ambient temperature can be monitored. One extra (LM-35) temp sensor can be connected to the COO board in order to read the inverter temperature and the ambient temperature around the CSS. An external 2.4GHz antenna ensures enough radio coverage in case where the PVAS is far away from CSS.

3.2 REMB

The REMB is the second part of the CSS and it is a board where all sensors are connected as it is shown in Fig. 2. As it is clear from the block diagram, a number of sensors like current and temperature are separated from the REMB. This is mainly because the measurement points (e.g. inverter voltage, current, or temperature) may be located far from REMB. The connection of sensors to the REMB is done through shielded stereo audio cables.

The reason that the CSS is divided into two parts (COO board and REMB) is mainly because with this implementation, all dangerous high voltages and currents are kept isolated from COO board and REMB where user might have access when the system is on. Other reason is that interferences from high voltages and currents are kept away from the ETRX2-PA module and the PC. Connection between REMB and COO board is done with a CAT5 Ethernet cable via the RJ-45 ports. The power needed for the REMB to operate (+5V) is supplied directly from the COO board. In Fig. 3 the developed REMB is depicted.

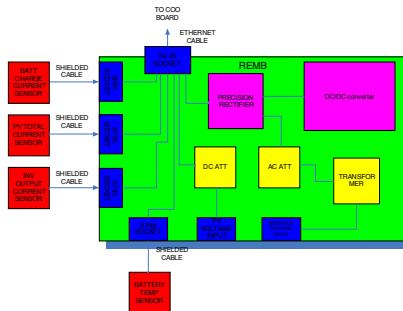


Fig. 2. REMB block diagram

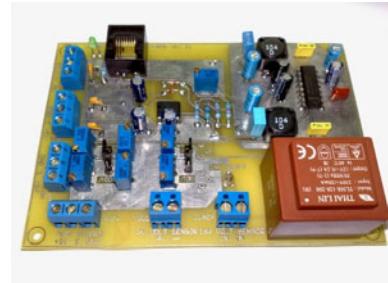


Fig. 3. The developed Remote Measurement Board (REMB)

3.3 Router

The router board provides a link between the ZEDs and the COO. The block diagram of the router is shown in Fig. 4. As already described previously, the router serves all end devices as a data sink to the COO board.

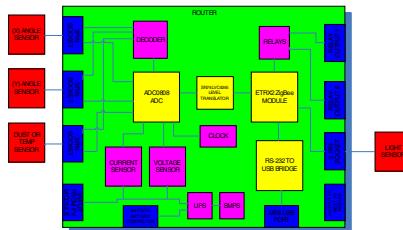


Fig. 4. Router block diagram

The Router has extra functionality compared to end devices. Except current, voltage and temperature, the router can monitor the air dust concentration and also the angle of the PV array in case of a tracking system is used. In addition, the router is equipped with two relay controlled outputs that can be used to control the motors, servos or the actuators of a tracking system and also a set of hall-effect angle sensors that are used to provide feedback to the tracking controller. The router board can also broadcast the sun illumination levels by simply connecting a Si photodiode or an

off-the-shelf pyranometer to one of its inputs. To avoid the use of batteries in the router board as well as in ZEDs and in order to operate at low PV panel output voltages (mainly at low insolation levels), a specially designed Switch Mode Power Supply (SMPS) is used. This configuration allows the device to operate normally when the PV module voltage exceeds a three volts (3V) threshold. After tests and redesigns, the SMPS/voltage regulator solution became the most efficient and suitable compared to the conventional linear regulated power supplies.

3.4 ZED

An end device is the simplest device on this monitoring system. The parameters that are monitored are the PV modules output voltage, current, and temperature. Fig. 5 shows the block diagram of the end device.

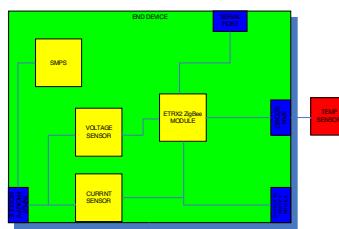


Fig. 5. Block diagram of a ZED

4 Measurements and Experiments

Two different experiments were conducted in order to verify the performance and reliability of the ZigBee PV monitoring system.

i. Bad connection experiment

In the first experiment, the case of a bad or corrupted connection has been simulated. In case N_o1, we disconnect the positive terminal of a PV module. This disconnection led to a zero current reading of the respective PV module (PV module n_o3) in the GUI application. The absence of current measurement but the presence of a 17V voltage reading leads to the conclusion that the PV module is disconnected from the DC bus. Such a state can be easily managed by a software application in order to enable an alarm. In case N_o2, we connected a PV module (PV module n_o4) in the DC bus but in that case corroded connectors used. This situation led to fluctuations of current and voltage readings in the GUI for the specific PV module confirming the bad connection.

ii. Low performance of a PV module

In this experiment, the case of a low performance caused by dirt or dust in the transparent surface of the module had been investigated. A mixture of fine sand and water had been sprayed to the glass surface of PV N_o4 simulating the red rain effect caused from the African dust, that is a very common phenomenon in the area of Crete. The PV monitor displayed a slight lower current reading in comparison to the mean value of the rest PV modules. The dust sensor discussed previously can alert for high

air dust concentration. Depending on the software extra reading can be extracted like power (W) and energy (Joules).

In Fig. 6 a number of graphs are shown representing the measurements corresponding to a specific day for each PV panel separately. The small differences between the graphs are mostly sensor errors. Another reason of these variations can be the fact that none of the PV module outputs the same amount of energy for a specified sun illumination.

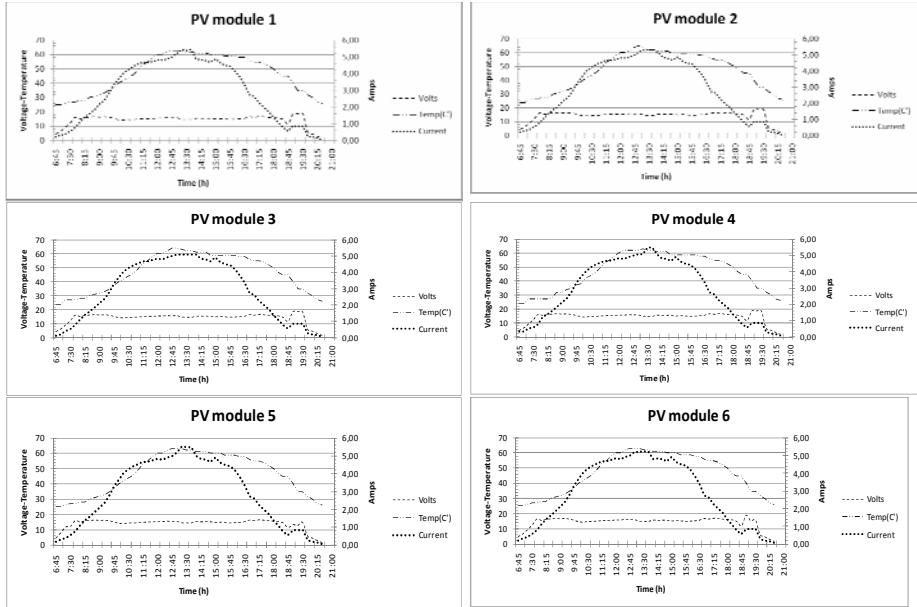


Fig. 6. PV current, voltage and temperature measurements of a specific day

5 Conclusions

In this paper, the prospect of using the ZigBee wireless technology in the field of PV supervising and monitoring was experimentally assessed. The system has been successfully tested on a low power PV system consisting of six (6) Photovoltaics modules (100Wp each) located at the main TEIC building (Chania/Greece). This kind of system can be installed in any kind of PV generation system regardless of size, as the ZigBee standard allows a very large volume of nodes (up to approximately 65,000) to be connected. The system presented here, provides accurate and also real time information about not only the overall PV plant behavior but also for any PV module alone.

Failures as well as disoperation of any component that consist the PV system can be identified immediately. As it is previously described, the system estimates and monitors the state of the PV plant through a strain forward process contrary to other systems that base their operation on indirectly methods such as complicated statistical

algorithms and comparisons of current performances with previous. In future works, connection of the system to the internet can be implemented in order for the user to observe the PV plant status from his personal computer. In case of no internet line availability, a GSM modem or other means of broadcast can be easily used.

Another important factor that can be characterized is the fact that here is no need of system inspection in the case of malfunctioning. Thus, specially trained personnel (maintenance operator) costs and also energy could be saved by the fact that the malfunctioning element is located very fast. Finally, the issue of integrating a ZigBee monitoring mote in each PV module in order to produce ready of-the-shelf ZigBee build in PV modules must be seriously investigated by the PV module manufacturers.

References

1. Gagliarducci, M., Lampasi, D.A., Podesta, L.: GSM-based monitoring and control of photovoltaic power generation. *Journal of Measurement* 40, 314–321 (2007)
2. Kalaitzakis, K., Koutroulis, E., Vlachos, V.: Development of a data acquisition system for remote monitoring of renewable energy systems. *Journal of Measurement* 34, 75–83 (2003)
3. Koutroulis, E., Kalaitzakis, K.: Development of an integrated data-acquisition system for renewable energy sources systems monitoring. *Journal of Renewable Energy* 28, 139–152 (2003)
4. Koizumi, H., Mizuno, T.: A novel micro controller for grid-connected photovoltaic systems. *IEEE Trans. Ind. Electron* 53(6), 1889–1897 (2006)
5. Gungor, V.C., Hancke, G.P.: Industrial Wireless Sensor Networks: Challenges, Design Principles, and Technical Approaches. *IEEE Transaction on Industrial Electronics* 56(10), 4258–4265 (2009)
6. Andrei, H., Dogaru-Ulieru, V., Chicco, G., Cepisca, C., Spertino, F.: Photovoltaic applications. *Journal of Materials Processing Technology* 181, 267–273 (2007)
7. Vergura, S., Acciani, G., Amoruso, V., Patrono, G.E., Vacca, F.: Descriptive and Inferential Statistics for Supervising and Monitoring the Operation of PV Plants. *IEEE Transaction on Industrial Electronics* 56(11), 4456–4464 (2009)
8. Drews, A., de Keizer, A.C., Beyer, H.G., Lorenz, E., Betcke, J., van Sark, W.G.J.H.M., et al.: Monitoring and remote failure detection of grid connected PV systems based on satellite observations. *Journal of solar energy* 81, 548–564 (2007)
9. Kay, R., Mattern, F.: The Design Space of Wireless Sensor Networks. *IEEE Wireless Communications* 11(6), 54–61 (2004)
10. Haenselmann, T.: Sensorsnetworks. GFDL Wireless Sensor Network Textbook (2006)
11. Chong, C., Kumar, S.P.: Sensor Networks: Evolution, Opportunities, and Challenges. *Proc. IEEE* 91(8), 1247–1256 (2003)
12. Culler, D., Estrin, D., Srivastava, M.: Overview of Sensor Networks. *Computer* 37(8), 41–49 (2004)
13. Jiang, Q., Manivannan, D.: Routing Protocols for Sensor Networks. In: Proc. 1st IEEE Consumer Comm. and Networking Conf., pp. 93–98. IEEE Press, Los Alamitos (2004)
14. Labiod, H., Afifi, H., Santis, C.: WI-FI, Bluetooth, ZigBee and Wimax. Springer, Netherland (2007)

A Linear Approach towards Modeling Human Behavior

Rui Antunes^{1,3}, Fernando V. Coito^{1,2}, and Hermínio Duarte-Ramos²

¹ UNINOVA

² Departamento de Engenharia Electrotécnica, Faculdade de Ciências e Tecnologia da
Universidade Nova de Lisboa
2829-516 Caparica, Portugal
{fjvc,hdr}@fct.unl.pt

³ Escola Superior de Tecnologia de Setúbal do Instituto Politécnico de Setúbal
2910-761 Estefanilha, Setúbal, Portugal
rui.antunes@estsetubal.ips.pt

Abstract. The human operator is, no doubt, the most complex and variable element of a Mechatronics system. On simpler manual control tasks, a linear model may be used to capture the human dynamics, however experiences on human operator response during pursuit manual tracking tasks, show that the dynamics of the human operator appear to depend on the specific task that the subject is asked to perform. This means that a unique truly human model cannot be completely achieved. Rather, a different set of models, each for a certain class of task, seems to be needed. This ongoing PhD work introduces several approaches on the human operator dynamic characteristic modeling and identification procedures, which may be useful for developing improved "humetronic" systems, i.e. human-machine systems which may be able to adapt themselves to the skill level of humans, aiming, with reduced effort, to achieve for best performance and safety.

Keywords: Human Dynamics Modeling, Identification, Human-Machine Interfaces, Manual Tracking Systems, Machine Adaptation.

1 Introduction

Nowadays we rely on many different mechatronics equipments and gadgets for carrying out our way of life, and we are well aware of the limitations with which machines can efficiently perform controlled manual tasks, because unfortunately these machines usually do not change regardless of the human operator's skill.

Current research aimed to improve performance and safety on man-machine systems is increasing. In an ordinary human-machine interface the human operator controls the machine by performing various manual control tasks. While performing the activity the operator is also learning it, and its skill is increasing. Hence, it seems natural the need for designing human-oriented machines which, in a smart way, may improve the assistance and performance with its human users.

1.1 Automatic Control versus Human-in-the-Loop (HIL) Control

Automatic control applications usually do not consider that the human factor inevitably belongs into the resulting closed-loop, and in an automatic control process human is still often considered remissness.

Human-in-the-loop (HIL) control is a starting approach to integrate the human model into the already assortment of electrical and mechanical factors involved in a closed-loop man-machine control design.

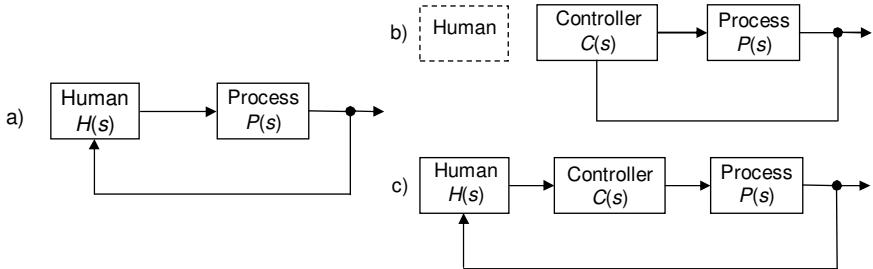


Fig. 1. a) Human-machine system *block diagram*. Comparison between *automatic* b) and *human-in-the-loop* control c).

1.2 The Human Adaptive Mechatronics (HAM) Assist Control Concept

An important goal today is to improve the design of next generation adaptive human-oriented machines, which will have the ability to intelligently cooperate with its human users. These machines should be able to actively adapt according to the skills of their operators, being capable to evaluate human's global performance. A human adaptive mechatronics assist control system [1], [2], [3] identifies the operator based on his acquired actions in the process, and an assist controller is then iteratively conceived to improve the operator's performance according to its skill.

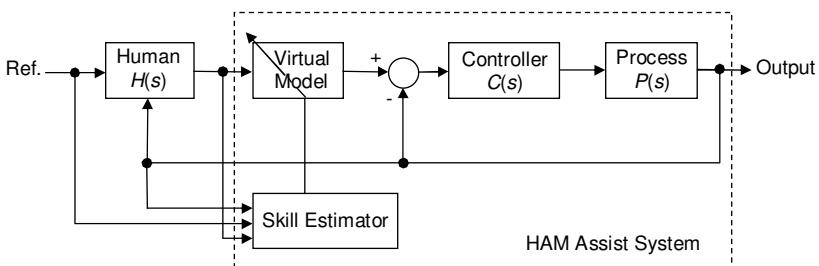


Fig. 2. Human adaptive mechatronics assist control *block diagram*

Fig. 2 shows a HAM assist control application example [4], where the machine dynamics is settled by tuning its virtual model from previously estimation of human's skill.

1.3 Modeling Human Behavior

Human models are important to design a machine system to be controlled by an operator. The model is not aimed to closely replicate the operator behavior, but to provide sufficient information for the design of closed-loop control system incorporating the human behavior. Thus, in spite the fact that humans are non-linear time-varying complex systems, linear models are used as a way to capture some of the relevant characteristics of the human operator. This allows the use of standard control design techniques, combined with supervisory controls approaches to tackle non-linearity and time variations.

To obtain linear models a number of experiments is required. Such experiments must be designed so as to minimize undesired effects like operator learning and prediction of the task or action signal saturation.

The first studies in human modeling were inspired by the demand for pilot models during the Second World War, and since that time there had been made many experiments, especially in pursuit and compensatory manual tracking tasks. One of the earliest studies of the human operator acting as a linear servomechanism was done in 1947 by Tustin, who proposed that there might be a linear law that could represent most of the operator's behavior. More recently new identification theories, methods and tools became available, striving human modeling to achieve significant progress.

In this paper several approaches are introduced on the human dynamic modeling and identification methods, and experimental procedures, for developing new human-oriented machines, which aim for best performance and safety, reducing also operator's effort. Although it must be stressed that whatever approach is taken to the description of human behavior, it only will capture a fraction of it, as human operator comprises the complex element in a man-machine system.

2 Contribution to Sustainability

Human-oriented machines help to create more sustainable services and products, without compromising human and environmental safety, as we are all challenged to produce and recycle more with fewer resources.

Essentially, three necessary phases are required for designing HAM systems:

- 1) Estimate human control characteristics;
- 2) Quantify the overall skill;
- 3) Design the assist-control human-machine system.

The first phase can be considered both as a signal processing and a system identification problem, which implies developing theoretical and experimental procedures for obtaining the human controller from measured response data. On simpler manual control tasks a linear model can capture relatively well the human's behavior¹. However, the model depends on the task and the process to be controlled, and simultaneously on many other factors, both personal and environmental, such as concentration, training, disturbances, comfort, workload, saturation, intermittency,

¹ The Crossover and the Proportional Derivative are generalized low order linear human models, well described in the literature.

fatigue, etc. The second phase for HAM design uses the experimental data obtained to quantify the overall operator skill, by comparing between an individual and its ideal response. Comparing ideal and identified parameters can also provide more accurate skill quantification. Next, according to the skill level obtained, an assist-control system (which is able to change the "amount of assistance") is finally implemented, based on an ideal model and on the acquired skill data.

In this work special focus is given on the first step of the Human Adaptive Mechatronics project, i.e. on the methods and procedures that can be derived for estimating the human control characteristics. We propose that the systemic notion of task dependence must be employed to model a human-machine system. New schemes for deriving task-dependent characteristics are explored through nonparametric and parametric identification methods.

Different approaches to the human identification problem are presented: we started by transient analysis (impulse and step response), obtained from pursuit manual tracking experiments. A special method to estimate the operator's delay-time was investigated. Structured Auto Regressive with eXogenous (ARX) input-output data parametric models were employed to obtain human-machine linear models from manual 1-D tracking experiences. Finally, the improved frequency analysis method is described in detail for modeling operator dynamics. An experimental setup that allowed the implementation of a human operator control task has also been developed, implemented and tested. The experimental procedures deployed for estimating the human-machine dynamics had an important role in this work, and were also described in detail.

3 State-of-the-Art / Related Literature

There have been in the past many studies covering the human operator dynamics modeling, especially within aeronautical engineering and flight control systems. Since the late 1940's that many experiments of human pilots performing target tracking tasks were studied. Important theories have been first introduced by McRuer et al., for manual tracking tasks using random references. A quasi-linear model was developed for describing human behavior during compensatory tracking tasks. In the 1960s and the 1970s modeling the human controller made great progresses [5]. The McRuer's et al. linear crossover model [6] and the optimal control model of Baron and Kleinman (applied for compensatory tracking) were proposed. The Baron's optimal control model was further developed by Tomizuka for preview tracking. In 1974 Shinnars proposed the use of ARMA (autoregressive moving-average) models, obtained from tracking tasks.

In the 1990s Kawato postulated that the human's control has feedback and feedforward structures, and through the process of learning he changes from feedback to feedforward (in the Feedback Error Learning model [7] the cerebellum acquires a model of the machine, as an inverse model in the feedforward path). Wolpert and Kawato improved the Feedback Error Learning model to a module selection and identification control (MOSAIC) [8], expanding the inverse model into a controller and the forward model into a predictor. Latest developments include using optimal control model (OCM) design techniques (by minimizing a certain performance index), assuming that the human operator performs the manual task in an optimal way. Particle swarm optimization (PSO) is a promising method for obtaining the

OCM parameters from experimental data. ARX models (linear approach) and adaptive-network-based fuzzy inference tools have been recently studied. A new hybrid fuzzy-ARX modeling method [9] has also been recently developed for predicting the human operator control actions.

4 Research Contribution and Innovation

This work introduces several methodologies on the human-machine dynamic modeling and identification procedures, which can be applied on the development of improved HAM systems. In general, for execution of complicated tasks the human response does not follow linear behavior. However, for simpler servo/regulator control tasks, a linear model can be employed to capture most of the operator's dynamic characteristics.

Linear models can be obtained by physical (mathematical) modeling, and through system identification techniques based on measured data. The first approach is rather involved and time consuming, lying far beyond the objectives in hand. Hence, this work is focused on the system identification problem, i.e., on how to estimate a human-machine system model from the observed input/output data. Parametric identification methods are techniques to estimate parameters for pre-defined model structures, by finding through numerical search the parameters values that minimize the differences between the model's output and the measured data. ARX parametric identification methods were used in this work to model the human-machine dynamics. Nonparametric identification methods allow obtaining the operator behavior without the need for a pre-defined parameterized model structure. We have employed also nonparametric methods in this work, by analyzing the human operator transient and frequency responses on manual tracking tasks.

4.1 ARX Model Estimation

Auto-Regressive with eXogenous terms (ARX) models are well suited for modeling linear systems, due to its high potential and simplicity. The most common ARX structure is the linear difference equation:

$$y(t) + a_1 y(t-1) + \dots + a_{na} y(t-na) = b_1 u(t-nk) + \dots + b_{nb} u(t-nk-nb+1). \quad (1)$$

The present output $y(t)$ is related to a number of past inputs $u(t-k)$ and past outputs $y(t-k)$. na equals the number of poles and nb the number of zeros. nk is the time delay inherent to the system. Typical estimation methods used for obtaining a and b parameters are the least-squares, the Kalman filter, and the instrumental variable method.

4.2 Transient Response Analysis

The dynamic properties of a linear model can be investigated by analyzing its transient response. For example, system time delays, time constants and static gain can be computed through the obtained impulse and step responses. Also, the right

previous selection of the delay nk in the ARX model structure may be crucial for obtaining good identification results. Finally, the identification of low order linear models can also be easily confirmed from transient (step and impulse) analysis experiments.

4.3 Time-Delay Estimation

Time-delay [10] is an important human factor that affects operator performance. Although human is considered a nonlinear system, many studies (such as proposed by Ragazzini) support that in a simple motor task the operator behavior can be sufficiently identified by a linear model plus a finite time-delay (as a result of the neuromuscular and central nervous latencies, and also due to other human dependent and environmental factors). The time-delay parameter is inherent to the ARX method, through the sampling shift. Therefore, human operator time-delay estimation can be further confirmed throughout the ARX identification procedure.

4.4 The Improved Frequency Analysis Method

Another nonparametric identification method is based on the frequency response, i.e. on how a linear dynamic model would react to certain sinusoidal inputs. In a LTI system, if we let the input $u(t)$ be a sinusoid of a certain frequency, then the output $y(t)$ will also be a sinusoid of the same frequency. However, the amplitude and the phase may be different. If we consider a one-dimensional normalized input signal of duration T , to be tracked, $x(t)$, built from a sum of N sinusoids at fixed multiple frequencies, the correspondent output $y(t)$ and the I/O response for each frequency may be obtained through the following diagram:

$$x(t) = x_0 + \sum_{k=1}^N a_k \sin(\omega_k t) \Rightarrow y(t) = y_0 + \sum_{k=1}^N b_k \sin(\omega_k t + \varphi_k) . \quad (2, 3)$$

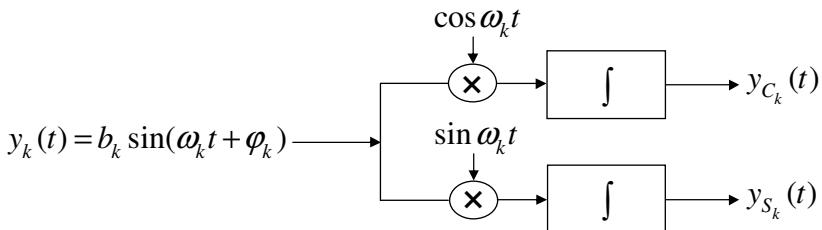


Fig. 3. Frequency analysis *block diagram* for each k -multiple frequency

By performing the integration along time $T = \frac{k2\pi}{\omega}$ (a multiple of the sinusoid period,), results in:

$$y_{C_k}(T) = \int_0^T b_k \sin(\omega_k t + \varphi_k) \cos \omega_k t dt \Leftrightarrow y_{C_k}(T) = \frac{b_k T}{2} \sin \varphi_k \quad (4.5)$$

$$y_{S_k}(T) = \int_0^T b_k \sin(\omega_k t + \varphi_k) \sin \omega_k t dt \Leftrightarrow y_{S_k}(T) = \frac{b_k T}{2} \cos \varphi_k \quad (6.7) \quad (8)$$

$$b_k = \frac{2}{T} \sqrt{y_{C_k}^2(T) + y_{S_k}^2(T)} \quad \text{and} \quad \varphi_k = \arctan \left(\frac{y_{C_k}(T)}{y_{S_k}(T)} \right) \quad K_0 = \frac{y_0}{x_0} .$$

which corresponds to the resulting human-machine closed-loop frequency response and static gain K_0 (for a previous input offset x_0). From the closed-loop experimental data, an open-loop human-machine model can be obtained, by inverse manipulation.

5 Discussion of Results and Critical View

This section presents obtained results from the experiments that were conducted for estimating human-machine models, in 1-D pursuit manual tracking tasks.

5.1 Experimental Procedures

For the transient analysis, two sets of five tracking samples (each lasted 2 minutes, and with at least 6 steps/impulses) were performed for a same operator with no history of neurological disease. To ensure human memorization or fatigue does not influence results, a minimum 15 minute rest was imposed between trials. The impulse and step inputs were visually recognized through large led indicators. In concerning ARX modeling, a set of four tracking tasks were conducted, with $T=120$ seconds duration each, comprising a white noise low-pass ($H(s)$) filtered input signal.

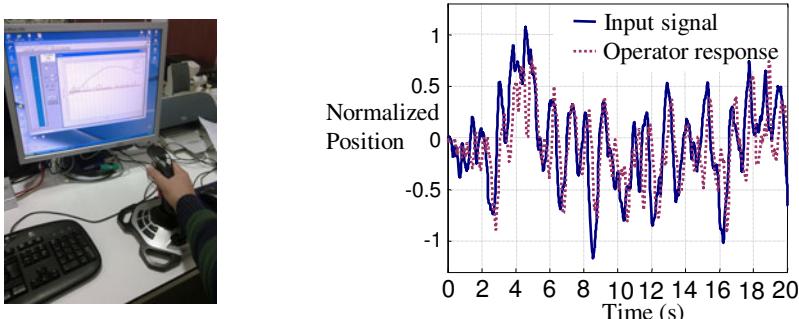


Fig. 4. A manual tracking time-trial using Logitech's Extreme 3D Pro. 8-bit analog Joystick

$$H(s) = \frac{25}{s^2 + 7.0711s + 25} . \quad (9)$$

To obtain a human-machine linear model from the experimental data, a procedure was proposed. In system identification the adequate choice of the sampling rate is important. Sampling too fast leads to poor modeling. However, the use of an incorrect value for pure time-delay can mask system behavior. So the procedure to choose both time-delay and sample rate is as follows:

1. The signals are oversampled, at 100 Hz.
2. The first 0.5 seconds of each test are discarded because it corresponds to the low-pass filter initial transient.
3. For each choice for time-delay the signals are decimated before an ARX model is computed. From the point of view of discrete time models this corresponds to non-integer delay.
4. Taking the quadratic deviation between the true output and its estimated value as a performance index, the best combination for (time-delay, sampling rate, model structure) is chosen.
5. Continuous time equivalents of the time-delay and ARX model are separately obtained, and combined into a single continuous time model.

Frequency analysis experimental procedures and results fall beyond the size of this paper, and were already described in detail in previous work [11], [12].

5.2 Experimental Results

Obtained step and impulse response data showed that pure time-delay is unrelated with the number of trials (or human experience). Hence, for the identification of the operator's behavior a simple compensation data shift can be included in the correspondent model. Saturation and nonlinearity characteristics can also be found at the maximum limits, and near the joystick's origin, and the impulse tests were made with a 0.2 offset, to avoid the mechanical nonlinearities near the origin.

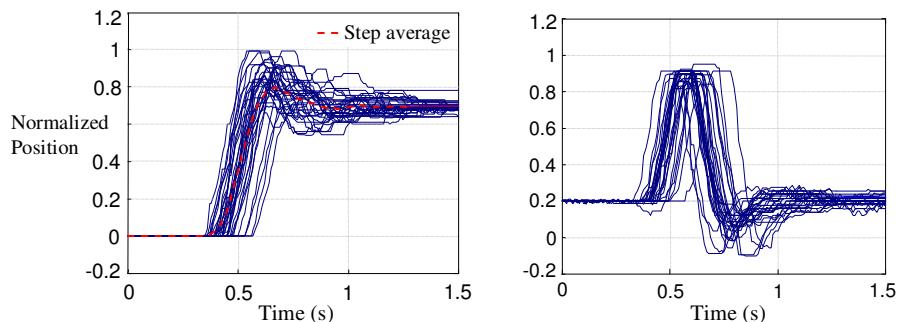


Fig. 5. Step (left) and impulse (right) responses, at 100Hz sampling rate

Table 1. Pure time-delay (*in seconds*), on five manual tracking step and impulse tasks

Impulse				delay (s)						
1 - 10	0.43	0.37	0.36	0.39	0.41	0.46	0.39	0.42	0.38	0.51
11 - 20	0.57	0.43	0.41	0.40	0.46	0.45	0.44	0.41	0.42	0.58
21 - 30	0.44	0.41	0.46	0.46	0.44	0.49	0.54	0.39	-	-
Step				delay (s)						
1 - 10	0.43	0.42	0.45	0.45	0.46	0.4	0.4	0.42	0.42	0.42
11 - 20	0.41	0.39	0.44	0.46	0.47	0.44	0.42	0.37	0.43	0.51
21 - 30	0.34	0.52	0.5	0.43	0.4	0.58	0.39	0.45	0.49	0.41

As for ARX modeling, the best fit corresponds to the model:

$$Y(s) = e^{-0.28s} M(s)$$

$$M(s) = \frac{61.42s^7 - 1.58e004s^6 - 3.483e005s^5 - 1.785e006s^4}{s^9 + 116s^8 + 5730s^7 + 1.691e005s^6 + 3.198e006s^5 + 4.197e007s^4} \cdot \frac{+3.243e007s^3 + 5.992e008s^2 + 3.584e009s + 4.209e009}{+3.656e008s^3 + 2.132e009s^2 + 7.016e009s + 6.316e009} \quad (10,11)$$

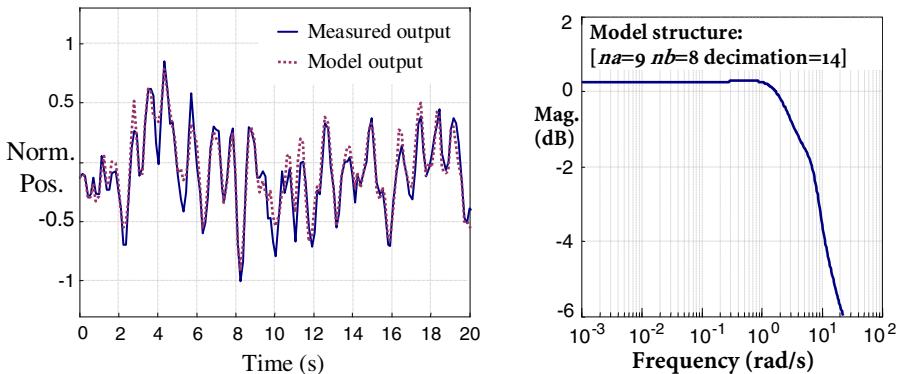


Fig. 6. ARX-980 model validation (*left*) on a manual tracking sample (*first 20 seconds*). Magnitude Bode plot for the human-machine ARX-980 model structure (*right*).

The human-machine estimated time-delay is 280 ms, while the average time-delay obtained from the impulse and step analysis was, respectively, 440 ms and 437 ms.

Although impulse response analysis is a simple and fast open-loop experimental method, it has some disadvantages, due to augmented saturation, nonlinearities and human fatigue. Step response analysis confirmed the same time-delay values that were obtained in the impulse response experiments. These methods, along with ARX model estimation are still limited modeling techniques, which need additional validation. Frequency analysis can be used to obtain an accurate open-loop LTI model, from a sum of sinusoidal input signals. It is a closed-loop experimental method. Hence, to capture the relevant human-machine dynamic characteristics, the obtained models need to be converted to open-loop ones.

6 Conclusions and Further Work

This work investigated the linear modeling methods potential for identifying the human-machine controller characteristics, as required for the design of Human Adaptive Mechatronics (HAM) systems.

An experimental human-machine LabVIEW interface setup has been developed and implemented for collecting data from human operators, as they performed pursuit manual tracking tasks with an analog joystick. As confirmed in previous work, these obtained models depend on the type and shape of the input tracking signal (the same occurs with the estimated time-delay). It was also verified that the time-delay characteristic (for a same individual) is almost constant regardless the skill level or experience. Hence, pure time-delay can be detached for identification purposes.

For future work, and to be even more applicable, the linear identification methods described should be also accomplished from real-time collected task execution data.

References

1. Harashima, F., Suzuki, S.: Human Adaptive Mechatronics - Interaction and Intelligence. In: 9th IEEE Int. Workshop on Advanced Motion Control, Istanbul, pp. 1–8 (2006)
2. Habib, M.: Human Adaptive and Friendly Mechatronics (HAFM). In: Proceedings of IEEE Int. Conf. on Mechatronics and Automation, Takamatsu, pp. 61–65 (2008)
3. Kado, Y., Pan, Y., Furuta, K.: Control System for Skill Acquisition – Balancing Pendulum based on Human Adaptive Mechatronics. In: IEEE International Conference on Systems, Man, and Cybernetics, Taipei, Taiwan, pp. 4040–4045 (2006)
4. Suzuki, S., Harashima, F.: Assist Control and its Tuning Method for Haptic System. In: 9th IEEE Int. Works. on Advanced Motion Control, Istanbul, Turkey, pp. 374–379 (2006)
5. Gaines, B.: Linear and Nonlinear Models of the Human Controller. International Journal of Man-Machine Studies 1, 333–360 (1969)
6. McRuer, D., Graham, D., Krendel, E., Reisener, W.: Human Pilot Dynamics in Compensatory Systems. Technical report n.º AFFDL-TR-65-15, Air Force Flight Dynamics Laboratory, Wright-Patterson AFB. Ohio (1965)
7. Ito, M.: Internal model visualized. Nature 403, 153–154 (2000)
8. Suzuki, S., Watanabe, Y., Igarashi, H., Hidaka, K.: Human skill elucidation based on gaze analysis for dynamic manipulation. In: IEEE International Conference on Systems, Man and Cybernetics, Montreal, pp. 2989–2994 (2007)
9. Celik, O., Ertugrul, S.: Predictive human operator model to be utilized as a controller using linear, neuro-fuzzy and fuzzy-ARX modeling techniques. In: Engineering Applications of Artificial Intelligence, vol. 23, pp. 595–603. Elsevier (2010)
10. Boer, E., Kenyon, R.: Estimation of Time-Varying Delay Time in Nonstationary Linear Systems: An Approach to Monitor Human Operator Adaptation in Manual Tracking Tasks. IEEE Transactions on Systems Man and Cybernetics 28(1), 89–99 (1998)
11. Antunes, R., Coito, F., Duarte-Ramos, H.: Using Human Dynamics to Improve Operator Performance. In: Camarinha-Matos, L.M., Pereira, P., Ribeiro, L. (eds.) DoCEIS 2010. IFIP Advances in Information and Communication Technology, vol. 314, pp. 393–400. Springer, Heidelberg (2010)
12. Antunes, R., Coito, F., Duarte-Ramos, H.: Human-Machine Control Model Approach to Enhance Operator Skills. In: Proceedings of IEEE International Conference on Mechanical and Electrical Technology, Singapore, pp. 403–407 (2010)

Nonlinear-Fuzzy Based Design Actuator Fault Diagnosis for the Satellite Attitude Control System

Alireza Mirzaee¹ and Ahmad Foruzantabar²

¹ Highgraduate in Control Engineering,

Electronic Department of Islamic Azad University-Dariun Branch, Shiraz-Iran

Mirzaee@gmail.com

² Phd Student in Control Engineering,

Electronic Department of Fars Science and Research Branch Islamic Azad University,

Marvdasht-Iran

A.foruzantabar@srbiau.ac.ir

Abstract. The objective of this paper is to develop a hybrid scheme (nonlinear observer and fuzzy decision making) for fault detection and isolation (FDI) in the reaction wheels of a satellite. Specifically, the goal is to decide whether a bus voltage fault, a current fault or a temperature fault has occurred in one of the three reaction wheels and further to localize which fault has occurred. In order to achieve these objectives, high fidelity model is used to exhibit the dynamics of the wheels on each of the three axes independently. First using the dynamic equations, nonlinear observer is designed, and then comparing estimated and actual states, residual signals are generated. These two input signals comprise the decision making unit. Design of the decision making unit using fuzzy reasoning is implemented. The effectiveness of this nonlinear-fuzzy based FDI scheme is investigated and a comparative study is conducted with the performance of a generalized observer-based scheme.

Keywords: Fuzzy decision making, qualitative reasoning, nonlinear observer, fault detection and isolation, fault diagnosis, reaction wheel, satellite, Takagi-Sugeno.

1 Introduction

Attitude control is an important basic function for most spacecrafts especially for satellites. It has been widely studied since late 1950s [1]. The attitude control subsystem stabilizes the spacecraft and orients it in the desired set point position in a short time despite the presence of external disturbance torques. The control torques could be formed from a combination of momentum wheels, reaction wheels, control moment gyros, thrusters or magnetic torquers. The main actuators for satellite attitude control systems are reaction wheels.

A high fidelity mathematical model of a reaction wheel [2] is discussed briefly in Section 2. Normally, there are three types of faults in a wheel that require special attention. The first is the bus voltage fault. The bus voltage should be sufficiently high to avoid elimination of the voltage headroom. Low bus voltage will result in reduced torque capacity and consequently cause serious instability of the satellite attitude. The same effect will appear when the motor current loss occurs in the wheel this leads to a

loss of power and therefore the wheel cannot supply enough reaction torque to achieve a proper set point change of the attitude. Finally, the temperature change is the third source of fault. The temperature is highly related to the viscous friction, which is the main friction factor of the wheel. The temperature fault will cause the wheel to operate in an abnormal condition. Model-based methods are a number of methods based on intelligent and learning-based strategies [3], [4], [5]. These methods make use of the large amount of process history data. Fuzzy logic technique has been investigated as powerful decision making tool as they can be used as supervisory schemes to make the fault analysis decisions [6], [7]. The basic idea behind the model based observer approaches is to estimate the states of the system by using either Linear or Nonlinear observers. The state estimation error is served as the residual. The advantage of using the observer is the flexibility to select its gains that leads to a rich variety of FDI schemes [8], [9].

In this paper, a nonlinear observer is employed for the reaction wheel of each axis so as to observe the estimated angular velocity and the motor current from each wheel. Among these estimated signals, one will be able to identify the existence and isolation of faults in the system. The outline of the remainder of this paper is as follows. In Section 2, contribution to sustainability is defined. Section 3, a brief review of the attitude control system and model of the reaction wheel will be given. Section 4 presents results for a nonlinear observer-based scheme used for fault detection. Section 5 presents results for a fuzzy-based scheme used for fault isolation. In Section 6, by combining two previous sections a nonlinear-fuzzy based FDI scheme will be developed. A comparative study between nonlinear-fuzzy and linear observer based FDI scheme will be conducted in Section 7. These results will serve as benchmark data for comparison with the proposed FDI scheme. These comparative results will demonstrate the advantages of nonlinear-fuzzy based scheme developed in this paper.

2 Contribution to Sustainability

The whole observation and decision making process is done by accurate computer systems and this insures the ability of the system to survive without human interference. The proposed method used for attitude control system is innovative and the results of that are very satisfactory and suitable to implementation and test on the real reaction wheel.

3 Reaction Wheel Model

A reaction wheel consists of a rotating flywheel, typically suspended on ball bearings, and driven by an inertial brushless DC motor. Fig.1 provides the fundamental relationships for a high fidelity mathematical model of a reaction wheel system. There are five main sub-blocks in the diagram: motor torque control, speed limiter, EMF torque limiting, motor disturbances and bearing friction and disturbances. The reaction wheel applied in this paper is the ITHACO's standard Type A. Its typical parameter values used can be found in [3]. Operating in space a satellite experiences many types of external environmental disturbance torques. Four main disturbances we considered here are: gravitation torque, solar pressure torque, magnetic torque and aerodynamic torque. It is assumed that the maximum external disturbance torque is the sum of these four maximum torques:

$$\text{DIS} = \text{DIS}_{\text{gg}} + \text{DIS}_{\text{sp}} + \text{DIS}_{\text{mf}} + \text{DIS}_{\text{ad}} = 5.68 \times 10^{-5} \text{ N-m}$$

We assume that the external disturbance torque is a normally distributed random signal with zero mean and variance:

$$\text{DIS}^2 = (5.68 \times 10^{-5})^2 [2].$$

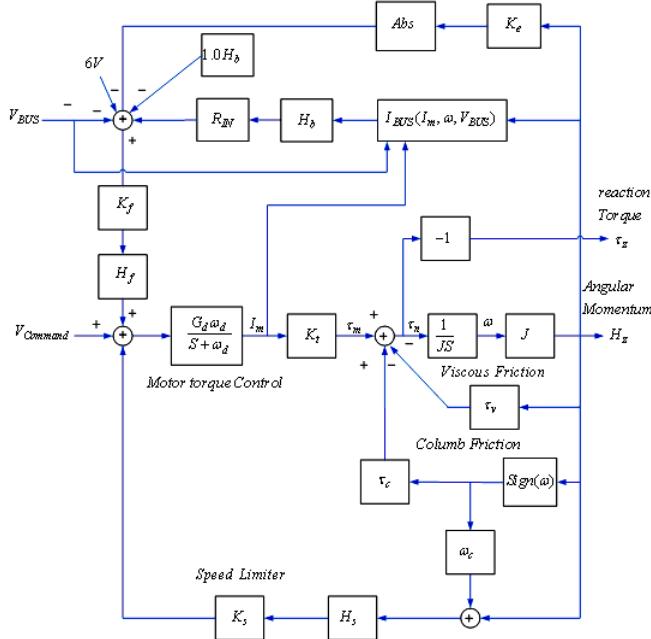


Fig. 1. High Fidelity Reaction Wheel Block Diagram

4 Nonlinear Observer-Based Fault Detection

FDI for nonlinear systems can be achieved by generating residuals using nonlinear observers. The nonlinear observer based residual generation problem can be formulated as follows; consider a nonlinear system given by (1).

$$\begin{cases} \dot{x} = f(x, u) \\ y = h(x, u) \end{cases} \quad (1)$$

Under the assumption of a smooth system [10], and assuming that H defined below is full rank, a simple form of a nonlinear estimator described by (2).

$$\begin{cases} \dot{\hat{x}} = f(\hat{x}, u) + L(\hat{x}, u)(y - \hat{y}), \\ y = h(\hat{x}, u) \end{cases} \quad H(\hat{x}, u) = \left. \frac{\partial \hat{f}}{\partial y} \right|_{\hat{x}, u} \quad (2)$$

Where H is a time varying observer gain matrix; the state estimation error equation and the output estimation error then are:

$$\begin{cases} \dot{\hat{e}}(t) = \left[\frac{\partial \hat{f}}{\partial y} - H(x, u) \frac{\partial h}{\partial x} \right]_{\hat{x}, u} e(t) \\ e(t) = y(t) - \hat{y}(t) = h(x, u) - h(\hat{x}, u) \end{cases} \quad (3)$$

The nonlinear system residual generator using a nonlinear observer is illustrated in Fig. 2.

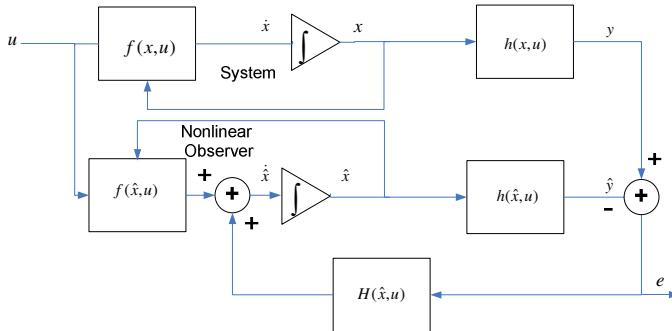


Fig. 2. Nonlinear Observer Residual Generator

The two variables ω and I_m in the reaction wheel model have been selected as state variables.

It is assumed that both these state variables are reachable.

Dynamic equations of the reaction wheel based on equation (4) are determined [11]. In this equation, f_1, f_2 are the functions for modeling motor disturbances, f_3 is output of EMF Torque Limiting block, f_4 is a sigmoidal function for modeling coulomb friction and f_5 represents the speed limiter block. n is the torque noise and r is the reference input or the torque command [11], [12].

$$\begin{aligned} \dot{\omega} &= \frac{1}{J} [f_1(\omega) + k_t I_m [f_2(\omega) + 1] - \tau_v \omega - \tau_c f_4(\omega) + n] \\ \dot{I}_m &= G_d \omega_d [f_3(\omega, I_m) - f_5(\omega)] - \omega_d I_m + G_d \omega_d r \end{aligned} \quad (4)$$

5 Fuzzy-Based Fault Isolation

After generation of residual Signals, the second step is decision making based on these residuals. At this step, given the characteristics of the two residuals, there should be three types of fault correctly isolated. Fuzzy logic as a powerful tool in inference and decision making based on linguistic terms and if-then rules used to isolate the three types of faults in the reaction wheel; Takagi-Sugeno, newer method of fuzzy inference than Mamdani Introduced in 1985 [13], is similar to the Mamdani method in many aspects. The first two parts of the fuzzy inference process, fuzzifying the inputs and applying the fuzzy operator, are exactly the same. The main difference between Mamdani and Sugeno is that the Sugeno output membership functions are either linear or constant. Based on this, Takagi-Sugeno is more suitable than other fuzzy implication methods. General Residual fuzzy evaluation scheme is illustrated in Fig. 3.

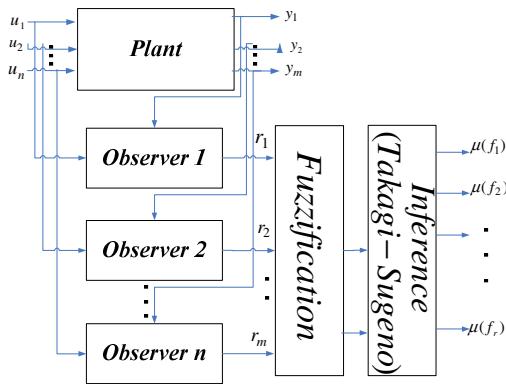


Fig. 3. Residual fuzzy evaluation

Where $u_i, i=1,2,\dots,n$ are the inputs, $y_i, i=1,2,\dots,m$ are the outputs, $r_i, i=1,2,\dots,m$ are the residuals, and $\mu(f_r)$ is the possibility of each hypothesized fault. Fig. 4 shows the fuzzification of the two residuals obtained in the previous section, the angular velocity and the motor current. Although the two residuals is not independent manner to the three types of faults, but using five membership functions this goal is attained. Each input in NB, NS, Z, PS and PB respectively indicating Negative Big, Negative Small, Zero, Positive Small and Positive Big. Intermediate membership functions are of Gaussian type and membership functions at the beginning and end, are respectively of S and Z type.

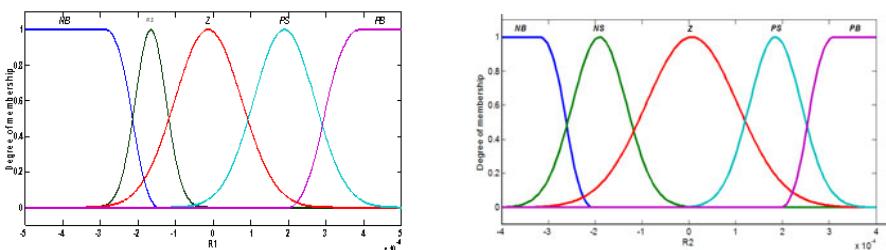


Fig. 4. Angular velocity (r_1), Motor current (r_2) membership Functions

For inference we can define:

$$\begin{cases} \mu_{r_1 \cap r_2}(f_r) = \mu_{r_1}(f_r) * \mu_{r_2}(f_r) \\ \mu_{r_1 \cup r_2}(f_r) = \mu_{r_1}(f_r) + \mu_{r_2}(f_r) - \mu_{r_1}(f_r)\mu_{r_2}(f_r) : x \in X \rightarrow [0,1] \end{cases} \quad (5)$$

In general, the number of rules derived from two inputs, to each output is 25. But according to the goal, and considering the simplicity and optimality of the design, based on the research done on the system and using the table 1 only 10 fuzzy rules are enough. Temp, Vbus and Im-Fault are outputs of the decision making unit. The output value is zero or one. Zero means a normal condition and one refers to a fault condition.

Table 1. Relationships between two residuals and three Faults

$r_1 \backslash r_2$	NB	NS	Z	PS	PB
NB	1	0	0	0	0
NS	1	0	0	0	0
Z	1	0	0	0	0
PS	1	0	0	0	0
PB	1	0	0	0	0

$r_1 \backslash r_2$	NB	NS	Z	PS	PB
NB	0	0	0	0	0
NS	0	0	0	0	1
Z	0	0	1	0	0
PS	0	0	0	0	1
PB	0	0	0	0	0

$r_1 \backslash r_2$	NB	NS	Z	PS	PB
NB	0	0	0	1	1
NS	0	0	0	0	0
Z	0	0	0	0	0
PS	0	0	0	0	0
PB	0	0	0	0	0

Voltage Bus Fault

Loss Current Fault

Temperature Fault

Combining these three tables, following rules can be obtained.

Rule 1 : if $r_1 = NB$ AND $r_2 = NB$ Then $VbusF = 1$ AND $ImF = 0$ AND $TempF = 0$

Rule 2 : if $r_1 = NS$ AND $r_2 = NB$ Then $VbusF = 1$ AND $ImF = 0$ AND $TempF = 0$

.

.

Rule 10 : if $r_1 = NB$ AND $r_2 = PB$ Then $VbusF = 0$ AND $ImF = 0$ AND $TempF = 1$

These rules are set in such a way that all three types of fault can be isolated into a wide range of operating points of attitude control system. In Takagi-Sugeno fuzzy decision making scheme, the weighted-average method has been used for defuzzification.

6 Simulation Results

When the normal value of the bus voltage (which is 8V) drops, the motor torque may be limited at high speeds due to the increasing back-EMF of the motor, and this eventually results in a reduced torque capacity of the wheel. When this value becomes too low (e.g. 3V), the attitude control system will malfunction and the attitude the spacecraft becomes unstable. Similarly, since the motor torque is directly related to the motor current through one constant parameter k_t , when some kind of motor current loss occurs in the reaction wheel, the motor torque will drop down accordingly. Therefore, the wheel can no longer supply enough motor torque to the attitude control system. When the current loss becomes significant, the controlled attitude angle will become unstable. It is well-known that viscous friction is present in the bearings due to the bearing lubricant. When the bearings are damaged seriously, this viscous friction becomes much larger than that when it is in normal conditions. Since the temperature of the wheel is strongly related to the viscous friction in the wheel, this suggests that an estimate of the working condition of the bearings is possible through monitoring wheel temperature.

6.1 Bus Voltage Fault

Fig. 5 shows a case study of the three faults. As seen, the bus voltage of the reaction wheel aligned on the X axis has dropped from the normal value of 8V to 5.5V in 200 seconds. After few seconds, bus voltage fault was detected. Hence, the bus voltage fault in the wheel of the X axis is detected and isolated correctly.

6.2 Current Loss Fault

Similarly, a current loss fault in the wheel of the X axis is properly detected and isolated as shown in Fig. 5. The current limiter signal of the wheel on the X axis has dropped from 1 to 0.40 in 100 seconds, implying that 60% of the motor current has been lost. Accordingly, the residual signal of the X axis has increased that output of fuzzy decision making unit is one after a short time delay and this indicated that the wheel on the X axis is faulty.

6.3 Temperature Fault

The simulations in Fig. 5 have illustrated that the proposed nonlinear-fuzzy scheme is also effective in detecting and isolating a temperature fault that has occurred in the wheel on the X axis at 100 seconds. Note that in each of the three cases, the residual curves for the other wheels will remain below their threshold.

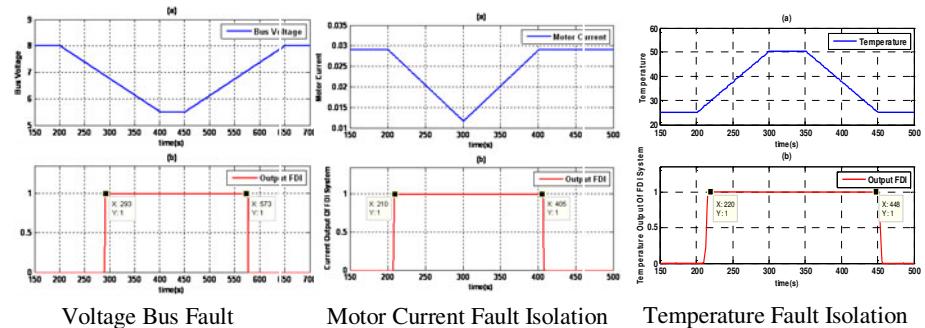


Fig. 5. Three Fault Isolation and their sensitivity to changes of voltage, current, temperature

7 Comparison of Linear and Nonlinear-Fuzzy Based FDI Schemes

The success of the two FDI schemes success highly depends on the modeling accuracy of the reaction wheel. As seen in Fig.1, reaction wheel is a highly nonlinear dynamic system. It contains many nonlinear elements and disturbances. For linear observer-based FDI scheme, one has to omit all these nonlinearities and disturbances effects and only use a linear model to observe the dynamics of the linearized reaction wheel model. In this paper, the objective is to design a fault diagnosis algorithm for the nonlinear model as the linear observer estimates are less likely to converge to states of a nonlinear model(not shown due to space limitations). With this in mind, a nonlinear observer is designed to improve the possible inefficiencies of a linear fault diagnosis observer. Choosing the nominal voltage bus that places the system in a less nonlinear area, this ensures that a linear observer is able to estimate to original nonlinear system.it could be shown that nevertheless it does not usually provide accurate information about the state of the system during the presence of faults. Thus it is not a good candidate for fault diagnosis in high nonlinear area.

8 Conclusion

This paper has developed and presented a nonlinear-fuzzy based scheme for fault detection and isolation in reaction wheels of a satellite. Faults considered are the bus voltage fault, the motor current loss fault and the temperature fault. First, we investigated a linear diagnosis scheme for fault detection and isolation. It is shown that this method has a capability for fault detection and isolation in some cases, but it is not reliable and useful under all operating conditions of the system. Since linear observer was not successful in detecting all types of faults, the performance of a nonlinear observer designed was shown that this observer can outperform the linear observer and detect all types of faults under different operating conditions of the system. Subsequently, a nonlinear-fuzzy based scheme is introduced to achieve better FDI performance of the reaction wheels. The proposed nonlinear-fuzzy based scheme consists of three networks applied to estimate the states of the wheels in three axes separately and to simplify the fault detection and isolation process. Through a comparative study between a linear and nonlinear-fuzzy scheme, it is demonstrated that the latter's performance is superior to the former's for FDI.

References

1. Wertz, J.R. (ed.): *Spacecraft Attitude Determination and Control*. Kluwer Academic Publishers, Dordrecht (1995)
2. Bialke, B.: High Fidelity Mathematical Modeling of Reaction Wheel Performance. *Advances in Astronautical Sciences* 98, 483–496 (1998)
3. Venkatasubramanian, V., Rengaswamy, R., Yin, K.: A review of process fault detection and diagnosis Part I: Quantitative model-based methods. *Computers and Chemical Engineering* (2002)
4. Isermann, R.: *Fault-Diagnosis Systems: An Introduction from Fault Detection to Fault Tolerance*. Springer, Berlin (2006)
5. Simani, S., Fantuzzi, C., Patton, R.J.: *Model-based Fault Diagnosis in Dynamic Systems Using Identification Techniques*, February 18. Springer, Heidelberg (2004)
6. Palade, V., Bocaniala, C.D., Jain, L. (eds.): *Computational Intelligence in Fault Diagnosis*. Springer, London (2006)
7. Soliman, A.A.: *The Application of Fuzzy Logic to the Diagnosis of Automotive Systems*. The Ohio State University (1997)
8. Frank, P.M.: Fault Diagnosis in dynamic systems using analytical and knowledge-based redundancy-A survey. *Automatica* 26, 459–474 (1990)
9. Frank, P.M.: Enhancement of Robustness on Observer-Based Fault Detection. *International Journal of Control*, 59(4), 955–983
10. Isidori, A.: *Nonlinear Control System: An Introduction*. Springer, New York (1989)
11. Azarnoush, H., Khorasani, K.: Fault Detection in Spacecraft Attitude Control System. *IEEE, Los Alamitos* (2007)
12. Li, Z.Q., Ma, L., Khorasani, K.: Fault Diagnosis of an Actuator in the Attitude Control Subsystem of a Satellite using Neural Networks. In: *Proceedings of International Joint Conference on Neural Networks*, Orlando, Florida, USA (2007)
13. Sugeno, M.: Industrial applications of fuzzy control. Elsevier Science Pub. Co., Amsterdam (1985)

Vergence Using GPU Cepstral Filtering

Luis Almeida^{1,2}, Paulo Menezes¹, and Jorge Dias¹

¹ Institute of Systems and Robotics,

Department of Electrical and Computer Engineering, University of Coimbra – Polo II,
3030-290 Coimbra, Portugal

{paulo,jorge}@isr.uc.pt

² Department of Informatics Engineering, Institute Polytechnic of Tomar
2300 Tomar, Portugal
laa@ipt.pt

Abstract. Vergence ability is an important visual behavior observed on living creatures when they use vision to interact with the environment. The notion of active observer is equally useful for robotic vision systems on tasks like object tracking, fixation and 3D environment structure recovery. Humanoid robotics are a potential playground for such behaviors. This paper describes the implementation of a real time binocular vergence behavior using cepstral filtering to estimate stereo disparities. By implementing the cepstral filter on a graphics processing unit (GPU) using Compute Unified Device Architecture (CUDA) we demonstrate that robust parallel algorithms that used to require dedicated hardware are now available on common computers. The overall system is implemented in the binocular vision system IMPEP (IMPEP Integrated Multimodal Perception Experimental Platform) to illustrate the system performance experimentally.

Keywords: Cepstrum, GPU, CUDA, vergence.

1 Introduction

Vergence ability is an important visual behavior observed on living creatures when they use vision to interact with the environment. In binocular systems, vergence is the process of adjusting the angle between the eyes (or cameras) so that they are directed towards the same world point. Robotic vision systems that rely on such behavior have proven to simplify tasks like object tracking, fixation, and 3D structure recovery. Verging onto an object can be performed by servoing directly from measurements made on the image. The mechanism consists of a discrete control loop driven by an algorithm that estimates single disparity from the two cameras. There are several methods to measure stereo disparities (features or area based correspondence, phase correlation based method, etc) and although some of them present better performance they were not used due to their computation requirements. Cepstral filtering is more immune to noise than most other approaches [1,2], but computing the Fast Fourier Transform (FFT) of images and inverse FFT presents some real-time challenges for the processing devices. This work describes the implementation of a real-time binocular vergence behavior using GPU cepstral filtering to estimate stereo

disparities. By implementing the real-time cepstral filter on a current graphics processing unit (GPU) using Compute Unified Device Architecture (CUDA) [4] we demonstrate that robust parallel algorithms can be used on common computers. The overall system is implemented in the binocular vision system IMPEP [23] (figure 1) to experimentally demonstrate the system performance. The main body of the cepstral algorithm, processed in parallel, consists of a 2-D FFT, a point transform (the log of the power spectrum), and the inverse 2-D FFT. The goal of the control strategy is to compensate the disparity between the cameras. Gaze holding behaviors and vergence processes are very useful for the emergent humanoid robotics area that aims to mimic humans. The following text presents the background for disparity estimation using cepstral filtering, a description of CUDA IMPEP implementation, experimental results and conclusions.

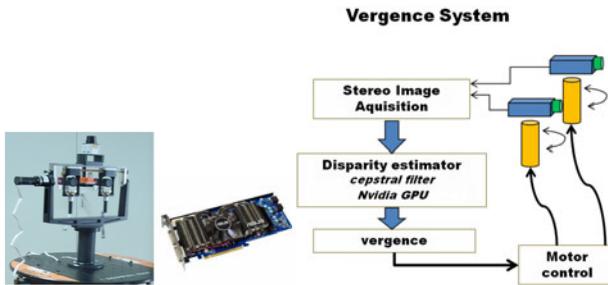


Fig. 1. Integrated Multimodal Perception Experimental Platform (IMPEP). The active perception head mounting hardware and motors were designed by the Perception on Purpose (POP - EC project number FP6-IST-2004-027268) team of the ISR/FCT-UC, and the sensor systems mounted at the Mobile Robotics Laboratory of the same institute, within the scope of the Bayesian Approach to Cognitive Systems project (BACS - EC project number FP6-IST-027140). On the right it is presented an overview of the IMPEP vergence system architecture and the NVIDIA GPU used for data parallel processing.

2 Contribution to Sustainability

Knowledge of the world allows the visual system to limit the amount of ambiguity and to greatly simplify visual computations. By demonstrating that computational power is available on computers at affordable costs we expect to contribute for the sustainability of computer vision complex tasks (intelligent surveillance systems, vision-guided autonomous vehicles, fingerprint/face/iris recognition, humanoid robotics, etc). The real-time cepstral filter implementation on a current graphics processing unit (GPU) using Compute Unified Device Architecture (CUDA) demonstrates that robust parallel algorithms can be used on common computers. By using the NVIDIA GPU multicore processors architecture and parallel programming we speed up the cepstral filtering algorithm more than sixteen times than on a CPU. The main body of the our GPU cepstral algorithm consists of a 2-D FFT, a point transform (the log of the power spectrum) and the inverse 2-D FFT. It takes only 0,43 ms to process an [256x256] image. The complete vergence control iterations cycle can be performed in 31ms (f=32,25Hz). The use GPU Cepstral Filtering to perform vergence on binocular head systems is, to our knowledge, an new contribution for the state-of-art.

3 Background and Related Work

Animals, especially predators, that have their eyes placed frontally can use information derived from the different projection of objects onto each retina to judge depth. By using two images of the same scene obtained from slightly different angles, it is possible to triangulate the distance to an object with a high degree of accuracy. For primates like ourselves the need for a vergence mechanism is obvious. Human eyes have non-uniform resolution, so we need a way to direct both foveas at the same world point so as to extract the greatest possible amount of information about it. The human brain has an extraordinary ability to extract depth information from stereo pairs, but only if the disparities fall within a limited range. Verging on surfaces usually constrains points near the fixation point to fall inside this range [2].

Binocular systems heads have been developed in recent decades. For example, VARMA head [12], MDOF head [13], Rochester [14], the "Richard the First" head [15] and the KTH robot head [16] were capable of mimicking human head motion. More recent robot heads include the POP head [23,24] used on the Bayesian Approach to Cognitive Systems project (IMPEP)[7], the LIRA-head [17], where acoustic and visual stimuli are exploited to drive the head gaze; the Yorick head [18], and the Medusa head [19] where high-accuracy calibration, gaze control, control of vergence or real-time speed tracking with log-polar images were successfully demonstrated.

In binocular camera systems, the vergence process has to adjust the angle between the cameras, by controlling the camera's pan angle, so that both sensors are directed at the same world point. The process must estimate the angle between the current direction of the non-dominant camera optical axis and the direction from the camera center to the desired direction (fixation point). The compensation angle is driven by continuously minimizing the binocular disparity. The IMPEP cameras do not have foveas. Even so, there are good reasons to have a low-level mechanism that maintains vergence. As Ballard and Olson argues [10,11], having a unique fixation point defines a coordinate system which is related as much to the object being observed as it is to the observer, and hence is a step in the direction of an object-centered coordinate system. Verging the eyes also provides an invariant that may be useful to higher level processes. It guarantees that the depth of at least one world point is known, even if we do not attempt stereo reconstruction in the usual sense. Additionally, by acquiring images that contain the focus of interest near the optical axis it is possible to avoid the effects due the camera lens radial distortion.

There are many different possible models for implementing vergence using disparity in the context of a robotic binocular system [2,3,6,11,12]. For example, by means of saliency detection or using stereo-matching techniques such as: phase correlation method like cepstral filtering, area based matching and feature-based matching. This work uses cepstral filtering to obtain a single disparity due their immunity to noise [1,2] and proves that the associated exponential calculus overhead (FFT) can be overcome by common parallel GPU's. Scharstein and Szeliski [21], and Brown [22], present thorough reviews of these techniques.

4 Visual Vergence Using Cepstral Disparity Filtering

A single disparity is estimated from the two cameras using the cepstral filtering. The cepstrum of a signal is the Fourier transform of the log of its power spectrum. Cepstral filter it is a known method of measuring auditory echo and it was introduced by Bogert [20]. The power spectrum of an audio signal with an echo present has a strong and easily identified component which is a direct measure of the echo period [1]. The binocular disparity measurement is obtained by applying of a non local filter (cepstral filter), followed by peak detection. Yeshurun and Schwartz [1,2] developed a method of using two-dimensional cepstrum as a disparity estimator. The first step of their method is to extract sample windows of size $h \times w$ from left and right images. The sample windows are then spliced together along one edge to produce an image $f(x,y)$ of size $h \times 2w$. Assuming that right and left images differ only by a shift, the spliced image may be thought as the original image at $(0,0)$ plus an echo at $(w+d_h, d_v)$, where d_h and d_v are the horizontal and vertical disparities. The periodic term in the log power spectrum of such signal will have fundamental frequencies of $w+d_h$ horizontally and d_v vertically. These are high frequencies relative to the window size. The image dependent term, by contrast will be composed of much lower frequencies, barring pathological images. Thus, as some authors [1] show, the cepstrum of the signal will usually have clear, isolated peaks at $(^+(w+d_h), ^+d_v)$.

The image $f(x,y)$ composed by the left and right images pairs can be mathematically represented as follow:

$$f(x, y) = s(x, y) * [\delta(x, y) + \delta(x - (W + d_h), y - d_v)] \quad (1)$$

Where $s(x,y)$ is the left image, $\delta(x,y)$ is the delta function, W the image width and $*$ operator represents two dimensional convolution. The Fourier transform of such image pair is

$$F(u, v) = S(u, v) \cdot (1 + e^{-j2\pi[(W + d_h)u + (d_v)v]}) \quad (2)$$

The power spectrum and the logarithm of equation (1), are:

$$|F(u, v)|^2 = |S(u, v) \cdot (1 + e^{-j2\pi[(W + d_h)u + (d_v)v]})|^2 \quad (3)$$

$$\log F(u, v) = \log S(u, v) + \log(1 + e^{-j2\pi[(W + d_h)u + (d_v)v]}) \quad (4)$$

and the Cepstral filter is the inverse Fourier transform of equation (4)

$$\begin{aligned} F^{-1}[\log F(u, v)] &= F^{-1}[\log S(u, v)] \\ &+ \sum_{n=1}^{\infty} (-1)^{n+1} \frac{\delta(x - n(W + d_h), y - nd_v)}{n} \end{aligned} \quad (5)$$

In the equation (5), the second term represents the prominent peak located in the output of Cepstral filter response. By determining these peak points positions it is possible to obtain disparity (figure 2).

4.1 Implementation on GPU Using CUDA

Our system uses the GeForce 9800 GTX+ with 128 cores and 512MB of dedicated memory to process the cepstral filter. The main body of the cepstral algorithm consists of a 2-D FFT, a point transform (the log of the power spectrum), and the inverse 2-D FFT.

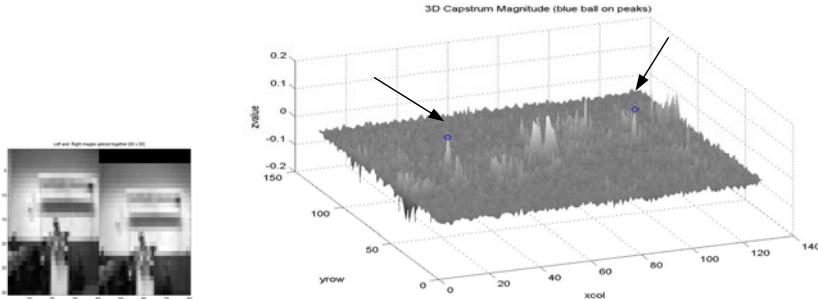


Fig. 2. Input subsampled spliced images [40x30], image pair with horizontal disparity=3 (left figure). Surface plot of the power spectrum of the cepstral filter (right figure). Peaks are visible at dominant global disparity location (marked with arrows).

For this 2D cepstrum algorithm we developed a GPU custom kernel to perform the point-wise absolute log in parallel using several threads, a GPU kernel to pad input data arrays with zeros (FFT requirement), GPU FFT transformations and all data allocation and data transfer procedures. A summarized global system algorithm loop is presented on figure 3. The 2D GPU FFT routines are implemented using CUFFT library [9], which are eight times faster than a CPU version using an optimized FFT and running on one core of a 2.4-GHz Core2 Quad Q6600 processor [8]. As the cepstral algorithm performs two FFT operations and the absolute log operation in parallel, the speedup is more than sixteen times faster than a CPU version. This multithreaded program is partitioned into blocks of threads that execute independently from each other, so that a GPU with more cores will automatically execute the program in less time than a GPU with fewer cores.

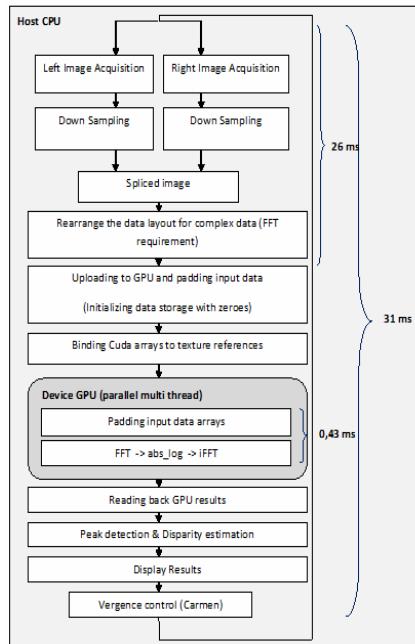


Fig. 3. Schematic block diagram of GPU cepstral filtering algorithm

5 Experiments

Experiment 1 – Image alignment

Figure 4 presents the real-time image alignment process frame sequence driven by the vergence control algorithm when an object is "instantly" positioned in front of the system. Both cameras changes alternate their angles to minimize the disparity. The performance measurements, according the schematic block diagram of figure 3, are shown on table 1.

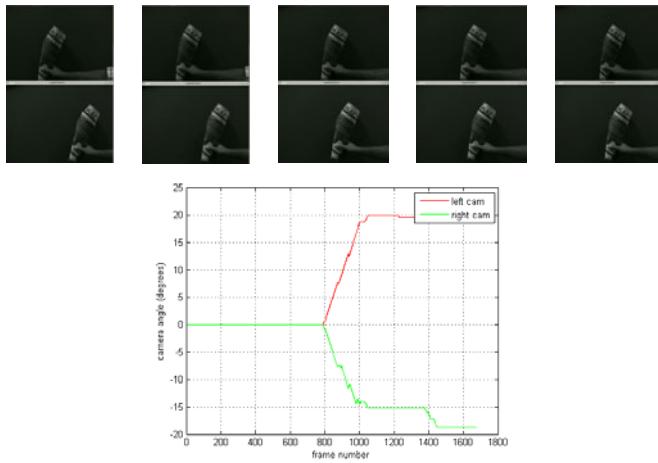


Fig. 4. – Real-time image alignment process frame sequence (each colum pair is an stereo pair). Below are the left camera angle values (red line) and right camera angle values (green) in degrees during the image alignment process.

Table 1. Processing time measurements

Task Set A	Processing Time	Task Set B	Processing Time
GPU (FFT abs log iFFT) [256x256]	0,43ms	GPU (FFT abs log iFFT) [256x256]	0,43ms
OpenCV image acquisition 2x[640x480] and preprocessing	26 ms	OpenCV image preloaded 2x[640x480] and preprocessing	3,2-4,5 ms
Complete iteration cycle with vergence control (f=32,25Hz)	31 ms	Complete iteration cycle without vergence control and image aquisition	6,9-9,1 ms (f=144,92Hz- 109,89Hz)

Experiment 2 – Image alignment with a dominant camera

We have also implemented an experiment where the left camera follows a color object (a ball) using CPU OpenCV camshift algorithm [5] and the right camera equally follows the object while trying to minimize the disparity using the GPU Cepstral Filtering (figure 5). By demonstrating this behavior we show that binocular heavy tracking algorithms can be applied to one only camera allowing CPU extra computational power for other tasks. Work on vergence controller should be carrying out to enable smooth movements.

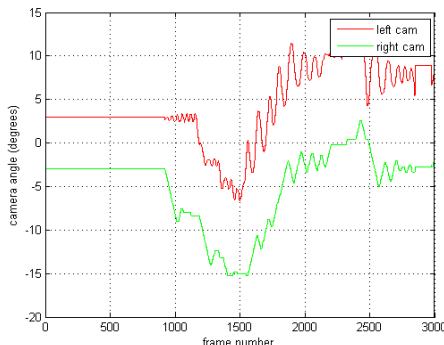


Fig. 5. Right camera follows left camera during a tracking task

6 Conclusions

By implementing the cepstral filter on a graphics processing unit (GPU) using Compute Unified Device Architecture (CUDA) we demonstrate that robust parallel algorithms that used to require dedicated hardware are now available on common computers for real time tasks. Using the GPU for low level tasks allows CPU extra computational power for other high level tasks. The cepstral filtering algorithm speed up is more than sixteen times than on a CPU and the use of GPU Cepstral Filtering to perform vergence on binocular head systems is, to our knowledge, an contribution for the state-of-art.

References

1. Yeshurun, Y., Schwartz, E.L.: Cepstral Filtering on a Columnar Image Architecture: A Fast Algorithm for Binocular Stereo Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 11(7), 759–767 (1989)
2. Coombs, D.: Real-time Gaze Holding in Binocular Robot Vision, PhD. Thesis, Department of Computer Science, University of Rochester (June 1992)
3. Kwon, K.-C., Lee, H.-S., Kim, N.: Hybrid Cepstral Filter for Rapid and Precise Vergence Control of Parallel Stereoscopic Camera. *Journal of the Research Institute for Computer and Information Communication* 12(3) (December 2004)
4. NVIDIA CUDA C ProgrammingGuide 3.1, NVIDIA (2010)
5. OpenCV (Open Source Computer Vision)
6. Taylor, J.R., Olson, T., Martin, W.N.: Accurate vergence control in complex scenes. In: Proc. Computer Vision and Pattern Recognition, Seattle, USA, pp. 540–545 (June 1994)
7. Ferreira, J.F., Lobo, J., Dias, J.: Bayesian Real-Time Perception Algorithms on GPU - Real-Time Implementation of Bayesian Models for Multimodal Perception Using CUDA. *Journal of Real-Time Image Processing* (February 26, 2010); Special Issue, Springer Berlin/Heidelberg, published online (ISSN: 1861-8219)
8. Garland, M., Le Grand, S., Nickolls, J., Anderson, J., Hardwick, J., Morton, S., Phillips, E., Zhang, Y., Volkov, V.: Parallel computing experiences with CUDA. *IEEE Micro.* 28(4), 13–27 (2008)

9. CUDA: CUFFT Library, NVIDIA Corp (2010)
10. Olson, T.J., Coombs, D.: Real-Time Vergence Control for Binocular Robots. IJCV 7(1), 67–89 (1991)
11. Ballard, D.H.: Reference Frames for Animate Vision. In: International Joint Conference on Artificial Intelligence. AAAI, Menlo Park (1989)
12. Dias, J., Paredes, C., Fonseca, I., Araujo, H., Batista, J., Almeida, A.T.: Simulating Pursuit with Machine Experiments with Robots and Artificial Vision. IEEE Transactions on Robotics and Automation 3(1), 1–18 (1998)
13. Batista, J., Dias, J., Araujo, H., de Almeida, A.: The ISR Multi-Degree of Freedom Active Vision Robot Head: Design and Calibration. In: SMART Program Workshop, Instituto Superior Técnico, Lisboa, Portugal, April 27–28 (1995)
14. Brown, C.M.: The Rochester robot, Tech. Rep. 257, Dept. Comp. Sci., Univ. Rochester, NY (1988)
15. Mowforth, P., Siebert, J., Jin, Z., Urquhart, C.: A head called Richard. In: Proceedings of the British Machine Vision Conference 1990, Oxford, UK, pp. 361–366 (1990)
16. Betsis, D., Lavest, J.: Kinematic calibration of the KTH head-eye system. In: ISRN KTH (1994)
17. Natale, L., Metta, G., Sandini, G.: Development of auditory-evoked reflexes: Visuo-acoustic cues integration in a binocular head. Robotics and Autonomous Systems 39, 87–106 (2002)
18. Eklundh, J.-O., Björkman, M.: Recognition of objects in the real world from a systems perspective. Kuenstliche Intelligenz 19(2), 12–17 (2005)
19. Bernardino, A., Santos Victor, J.: Binocular tracking: integrating perception and control. IEEE Transactions on Robotics & Automation 15(6), 1080–1094 (1999)
20. Bogert, B., Healy, M., Tukey, J.W.: The Quefrency Alanysis of Time Series for Echoes: Cepstrum, Pseudo-autocovariance, Cross-Cepstrum, and Saphe Cracking. In: Rosenblatt, M. (ed.) Proc. Symp. Time Series Analysis, pp. 209–243. John Wiley and Sons, Chichester (1963)
21. Scharstein, D., Szeliski, R.: A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms. International Journal of Computer Vision 47(1-3), 7–42 (2002)
22. Brown, M.Z., Burschka, D., Hager, G.D.: Advances in Computational Stereo. IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI) 25(8), 993–1008 (2003)
23. POP project (Perception on Purpose), number FP6-IST - 2004-027268,
<http://perception.inrialpes.fr/POP/>
24. Wilming, N., Wolfsteller, F., König, P., Caseiro, R., Xavier, J., Araújo, H.: Attention Models for Vergence Movements based on the JAMF Framework and the POPEYE Robot. VISSAPP 2, 429–435 (2009)

Motion Patterns: Signal Interpretation towards the Laban Movement Analysis Semantics

Luís Santos and Jorge Dias

Instituto de Sistemas e Robótica
Departamento de Engenharia Electrotécnica e de Computadores
Universidade de Coimbra, Portugal
3030-290 Pólo II
{luis,jorge}@isr.uc.pt

Abstract. This work studies the performance of different signal features regarding the qualitative meaning of Laban Movement Analysis semantics. Motion modeling is becoming a prominent scientific area, with research towards multiple applications. The theoretical representation of movements is a valuable tool when developing such models. One representation growing particular relevance in the community is Laban Movement Analysis (LMA). LMA is a movement descriptive language which was developed with underlying semantics. Divided in components, its qualities are mostly divided in binomial extreme states. One relevant issue to this problem is the interpretation of signal features into Laban semantics. There are multiple signal processing algorithms for feature generation, each providing different characteristics. We implemented some, covering a range of those measure categories. The results for method comparison are provided in terms of class separability of the LMA space state.

Keywords: Laban Movement Analysis, Motion Pattern, Signal Processing, Feature Generation.

1 Introduction

This paper sheds light on the interpretation of a human motion signal into a set of characteristics belonging to the Laban Movement Analysis semantics [1]. Despite the existence of multiple solutions for sensing human motion, this work is based on the study of body part trajectories, independent of the acquisition method. The objective is to apply multiple feature generation algorithms to segment the signals according to LMA theory, in order to find patterns and define the most prominent features in each of the descriptors defined in Labanotation [2]. This work can be seen as an important issue in human motion modeling, in the sense that feature generation strongly influences the model performance.

By definition, model is an abstract representation that reflects the characteristics of a given *entity*, either physical or conceptual. Thus, one issue of paramount importance is the establishment of the relation between sets of variables belonging to different abstraction levels.

The *entity* to be modeled sometimes has a theoretical representation/formalism that can be used as a basis for model development. In the specific case of human motion, LMA can be defined as a language to describe human motion in general and its application to human movement modeling is increasing [2],[4],[5]. LMA is divided in four¹ main components [3], each of them described through specific semantics that quantify and qualify different aspects of the human motion. Apart from mathematical sciences, LMA is a widely used tool in areas like physiotherapy, individual sports analysis, dancing.

We propose to study different feature generation/signal processing algorithms (e.g. Principal Component Analysis [6] or the analysis in the frequency domain [7]) to segment body part trajectories. The purpose of applying multiple segmentation techniques is to have a broad range of algorithms that provide different characteristics. Thus, this work provides a variety of features allowing finding suitable patterns that characterize each of the states of the Laban semantic space. We will evaluate the selected algorithms using a method based on Scatter Matrices to quantify the class separability, i.e. how each algorithm performs in terms of discretizing the variables in the LMA space-state.

This paper will be organized as follows: the next chapter will comment the contribution emerging from our work. Section 3 covers LMA theory, introducing a contextualization to Laban semantics. The subsequent section 4 will present the different segmentation algorithms, followed by the results in section 5. We will conclude in section 6 with the final remarks and future work.

2 Technological Innovation for Sustainability

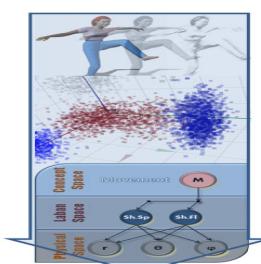


Fig. 1. The path from trajectory to the feature space to the classifier

During our research, we found residual contributions applying other techniques. It becomes even more noticeable when it comes to the use of Laban Movement Analysis as intermediate mid-level descriptor. The conducted research verified that most approaches use a set of theoretically defined features, rather than testing multiple methods towards the selection of good features relating to LMA semantics. This will help to improve motion/behavior model performance. Applicable to areas like surveillance, monitoring or physiotherapy, the performance improvement of such systems might have significant scientific impact with reflex on social and even economic sustainability.

A research on state of the art shows that most of features selected are very specific to the objectives of each study. Some work verses on general feature selection; however, most focuses on joint angle information and kinematics. During our research, we found residual contributions applying other techniques. It becomes even more noticeable when it comes to the use of Laban Movement Analysis as intermediate mid-level descriptor. The conducted research verified that most approaches use a set of theoretically defined features, rather than testing multiple methods towards the selection of good features relating to LMA semantics. This will help to improve motion/behavior model performance. Applicable to areas like surveillance, monitoring or physiotherapy, the performance improvement of such systems might have significant scientific impact with reflex on social and even economic sustainability.

¹ Laban theorists are not in unison regarding the number of main components. The two mainstreams divide themselves between four and five components respectively.

3 Interpreting with Laban Movement Analysis

Laban Movement Analysis has been described in previous works [4],[5], however it will be briefly introduced, in order to contextualize this work.

Developed in the early 20th Century by Rudolf Laban, Laban Movement Analysis has evolved throughout the years as a language to describe human motion, using a specific notation (Labanotation). Its semantic allows qualifying human motion in its different aspects, and introduces the fundamentals of our space-state definition.

Laban components are divided in two main groups, kinematic and non-kinematic. The kinematic components, body and space, deal with more quantitative aspects of the movement, and previous works [5],[8] demonstrate they are easily extracted. Consequently we found no value trying to apply those components the processing methods. The decision was to place emphasis on the qualitative (non-kinematic) components, Shape and Effort, as these pose a more relevant and interesting problem. Non-Kinematic components are described in theory by a rich and consequently complex semantic, thus constituting a very useful characterizing framework for human motion modeling. This work will not describe these two components in detail; rather it will make a very short overview and present the resulting space state. The Effort component is divided in 4 qualities lying between 2 extreme states. Each quality is associated to a cognitive process, a subject and lies between extreme states (see Table 1).

Table 1. Effort qualities, cognitive process, subject and space state

Quality	Cognitive process	Subject	Space State
Space	Attention	Spatial Orientation	[Direct, Indirect]
Weight	Intention	Impact	[Strong, Light]
Time	Decision	Urgency	[Sudden, Sustained]
Flow	Progression	How to keep going	[Free, Careful]

Bartenieff and Lewis [3] does not define Shape as a component of its own, but rather a set of qualities emerging from Body and Space components. Shape is also divided into two qualities, which are summarized in Table 2, defining the space-state.

Table 2. Shape qualities and correspondent space state

Quality	Space State
Flow	[Rising, Sinking] [Spreading, Enclosing]
Spatial	[Advancing, Retreating]

4 Feature Space

In the previous section, we have presented the space state of Laban Movement Analysis, which has only two states (binomial) for each quality. Its semantic carries meaningful qualitative characteristics which we seek to interpret using different signal processing algorithms.

The core of feature generation is to transform the available set of data features into another. If the transform is suitable, the transformation domain features may exhibit characteristics that yield a lot of meaningful information about the original signal.

In the feature generation area are domains which are more recurrent than others. Some algorithms aim data reduction, such as Principal Component Analysis and Single Value Decomposition [9], which belong to a class of methods known as Linear Discriminant Analysis. Within Nonlinear methods there are some focusing on the geometric characteristics of the signal in graph based approaches like Isometric Mapping. The Fourier Transform and others alike study the signal in the frequency domain. And there are a wide range of methods that study the signal regarding its derivative characteristics, the first, second and higher order moments methods.

4.1 Feature Generation Methods

Since the implementation of all methods is an intractable task, we selected a group of methods for this first approach, enough to cover the previously described domains in feature generation. The objective is to get a first evaluation on how each domain and correspondent methods behave in the task of discriminating the binomial LMA space state.

4.1.1 Karhunen-Loëve Transform

The computation of the Karhunen-Loëve (KL) transformation matrix will exploit the statistical information describing the data. The first assumption is that the data values have zero mean. The goal is to obtain mutually uncorrelated features in an effort to avoid information redundancies.

The method computes the data correlation matrix, which by its symmetric properties generates a set of mutually orthogonal eigenvectors V , known as the KL transform. As it turns out, KL has a number of other important properties, which provide different ways for its interpretation. One is the actually generated orthogonal eigenvectors, which encompass the principal directions of the spanned data, as well as the variance along each its directions. Thus we will use this information to represent trajectories in the resultant component space. We decided to use this information rather than the original purpose of the KL (re-project data in a dimensional space smaller the original), because data reduction methods are not optimized regarding class separability, and they do not assure that the principal components provide the best discriminatory properties.

KL transform, is a widely recognized technique, hence, more information on the method and its properties can be found in [6].

4.1.2 Local Linear Embedding

The starting point of this method is the assumption that the data points lie on a smooth manifold (hyper-surface). The main philosophy behind Local Linear Embedding [10] is to compute a low-dimensional representation of the data however preserving the local neighborhood information. The outcome of this algorithm attempts to reflect the geometric structure of the data. This algorithm can be resumed in its basic form with the following three steps:

(1) Select the nearest neighbors for each of the data points x_i , $i=1,2,\dots,n$. Some common techniques are Euclidean distances or the K-nearest neighbors.

(2) Compute the weights $W(i,j)$ that best reconstruct the point x_i from its neighbors minimizing the cost function

$$\arg \min E_W = \sum_{i=1}^n \left\| x_i - \sum_{j=1}^n W(i,j) x_j \right\|^2 \quad (1)$$

A typical weight function is

$$W(i,j) = \begin{cases} \exp \left(-\frac{\|x_i - x_j\|^2}{\sigma^2} \right), & \text{if points correspond to neighbors} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

where σ^2 is a user-defined parameter. The weights are constrained such that the rows of the weight matrix, i.e., the sum of the weights over all neighbors equals to 1.

(3) Use the weights obtained from the previous step to compute the corresponding points $y_i \in R^m$, $i=1,2,\dots,n$, to minimize the cost with respect to the unknown points $Y=\{y_i, i=1,2,\dots,n\}$

$$\arg \min E_Y = \sum_{i=1}^n \left\| y_i - \sum_j W(i,j) y_j \right\|^2 \quad (3)$$

This method explores the local linearity of the data and tries to predict each point through its neighbors using the least squares error criterion. Minimizing the cost regarding to the constraint given in (2) results in a solution that satisfies the following interesting properties: Scale, rotation and translation invariance.

Solving (3) for the unknown points y_i , $i=1,2,\dots,n$, is equivalent to:

- Performing the eigen-decomposition of the matrix $(I - W)^T (I - W)$.
- Discarding the eigenvector corresponding to the smallest eigenvalue.
- The remaining eigenvectors corresponding to the other eigenvalues yield the low-dimensional outputs y_i , $i=1,2,\dots,n-1$.

4.1.3 Discrete Fourier Transform

The Discrete Fourier Transform (DFT) [11] transforms a function into a sum of functions that represent it in the frequency domain. There is an assumption that the signal must be finite, which is accomplished in our case due to signal nature. The aim of this technique is to quantify how much of the signal lies in a determined frequency, i.e. to determine the dominant frequencies in a signal. For this work, we use the dominant frequencies and their coefficients to define the feature space state. We will not explain the theory, as this is probably one of the most well-known techniques around. However, we suggest the reader, if needed, to learn more or familiarize with method [11].

4.1.4 7 Moments of Hu

Under the scope of geometric moments, which are used to characterize data such as areas or information about orientation, we have the known 7 moments of Hu [10].

Within this class of methods, we have opted for Hu's moments because this technique intrinsically encompasses invariance to rotation, translation and scale. These are important properties because of the assumption that trajectory contours can be performed at different scales and orientations or space, depending on the physical structure of performer. The moments of Hu base themselves in the definition of central moments

$$\mu_{pq} = \iint (x - \bar{x})^p (y - \bar{y})^q f(x, y) dx dy \quad (4)$$

which are then normalized. We will not describe the mathematics of Hu's 7 moments, as they are somewhat cumbersome to this article and are readily available in [10] for the interested reader. An important remark is the statement that the first six moments are also invariant under the action of reflection, while the seventh moment changes signal. This property is interesting in the sense that it allows both left and right handed performers to be considered indifferent in terms of generated data. The values of these quantities can be quite different. In practice, to avoid precision problems, the logarithms of their absolute values are usually used as features.

4.2 Method Comparison and Evaluation

To establish a comparison criterion to evaluate the class separability capability of each method, we will use a method based on Scatter Matrix (SM) [10]. The reader might be familiar with the known Fisher Discriminant Ratio, which is a particular case of SM methods for 1 dimension and 2 classes. We selected SM due to the fact that other methods such as Divergence or Bhattacharyya Distance turn to be computationally demanding if a Gaussian assumption of data distribution is not employed. We should aim to select features leading to large between-class distance and small within-class variance in the feature vector space.

SM is built upon information related to the way feature vector samples are scattered in the l -dimensional space. The method defines the following matrices:

$$S_w = \sum_{i=1}^n P_i \Sigma_i \quad (5)$$

Which is known as *within class scatter matrix*, and Σ_i is the covariance matrix for class w_i and P_i is the a priori probability of class w_i , i.e. $P_i \simeq n_i/N$, where n_i is the number of samples of class w_i out of a total N samples. Then defining the *Between-class scatter matrix*

$$S_b = \sum_{i=1}^M P_i (\mu_i - \mu_0)(\mu_i - \mu_0)^T \quad (6)$$

where μ_0 is the global mean vector. The simplified computation for the *Mixture scatter matrix* turns out

$$S_m = S_w + S_b \quad (7)$$

with S_m the covariance matrix of the feature vector with respect to the global mean. Its *trace*² is the sum of variances of the features around they global mean. From these definitions we define the criterion as

$$J_1 = \frac{\text{trace}\{S_m\}}{\text{trace}\{S_w\}} \quad (8)$$

The ratio J_1 takes large values when samples are well clustered around their mean and the clusters are well separated.

5 Experimental Considerations

Most of the experimental process is an undergoing work as we aim to have a large enough database to encompass movements with all different LMA characteristics. We present our preliminary results using 5 different movements: Punch, Write, Waving Bye-Bye, Point and Lift. These movements have been hand labeled with different Effort Time and Effort Space characteristics. Punch, Point, Lift belong to *Direct* movements, whereas Write and Bye-Bye to *Indirect*. In the case of Effort Time, Punch and Point have been considered *Sudden* while the remaining three are considered *Sustained*. We have performed feature generation with all described techniques. Table 3 presents the separability ratio resulting from the application of the Scatter Matrix approach.

Table 3. The table shows the value of the separability ratio for the Effort Time and Effort Space qualities for each of the presented techniques

	PCA	DFT	LLE	Hu
Effort.Time	246,7	36,6	190,3	143,2
Effort.Space	210,2	29,1	229,6	134,3

6 Conclusions and Future Work

From the observation of the presented results, one concludes there is not a single perfect method for feature extraction. Different LMA qualities exhibit better separability performances for different methods. If we chose one method only, then we need to select one whose average performance is better. However if the computational cost of having different algorithms performing data processing is not an issue, then the choice must fall on the best method for the specific characteristic to be modeled. In the future there we will (it is an ongoing work) augment the database into a comprehensive set, which will encompass movements with all different LMA characteristics. The development of software to allow testing any desirable method is on the horizon, as well as doing efficiency vs. separability tests. Also, simple models for LMA classification should be done, for classification tests. The goal is to verify the true impact of the separability ratio vs. positive classification rate.

² Trace is defined to be the sum of the elements on the main diagonal.

References

1. Zhao, L.: Synthesis and Acquisition of Laban Movement Analysis Qualitative Parameters for Communicative Gestures. PhD Thesis, University of Pennsylvania (2002)
2. Chi, D., Costa, M., Zhao, L., Badler, N.: The emote model for effort and shape. In: Proceedings Annual Conference Series, Computer Graphics, SIGGRAPH 2000, pp. 173–182. ACM Press, New York (2000)
3. Bartenieff, I., Lewis, D.: Body Movement: Coping with the Environment. Gordon and Breach Science, New York (1980)
4. Rett, J.: Robot Human Interface Using Laban Movement Analysis Inside a Bayesian Framework. PhD Thesis, University of Coimbra (2009)
5. Rett, J., Santos, L., Dias, J.: Laban Movement Analysis using Multi-Ocular System. In: International Conference on Intelligent RObots and Systems, IROS (2008)
6. Jolliffe, I.: Principal Component Analysis, 2nd edn. Series in Statistics. Springer, NY (2002)
7. Broughton, S.A., Bryan, K.: Discrete Fourier analysis and Wavelets: Applications to Signal and Image Processing. Wiley, New York (2008)
8. Prado, J., Santos, L., Dias, J.: Horopter based Dynamic Background Segmentation applied to an Interactive Mobile Robot. In: 14th International Conference on Advanced Robotics, ICAR (2009)
9. Horn, R.A., Johnson, C.R.: Matrix Analysis. Cambridge University Press, Cambridge (1985)
10. Theodoridis, S., Koutroumbas, K.: Pattern Recognition, 4th edn. Elsevier, Amsterdam (2009)
11. Oppenheim, A.V., Schafer, R., Buck, J.: Discrete-time signal processing. Prentice Hall, Upper Saddle River (1999)

ARMA Modelling of Sleep Spindles

João Caldas da Costa¹, Manuel Duarte Ortigueira², and Arnaldo Batista²

¹ Department of Systems and Informatics, EST, IPS, Setubal, Portugal

² UNINOVA and Department of Electrical Engineering, University Nova, Lisbon, Portugal

Abstract. Differences in EEG sleep spindles constitute a promising indicator of neurodegenerative disorders. In this paper an ARMA modelling to sleep spindles is proposed and tested. The primary objective is to distinguish, via poles and zeros location, between regular, elderly and dementia subjects. In order to achieve this goal, a model validation has been done.

Keywords: sleep spindles, ARMA, EEG.

1 Introduction

Sleep spindles are particular EEG patterns which occur during the sleep cycle with center frequency in the band 11.5 to 15 Hz. They are used as one of the features to classify the sleep stages [1]. Sleep spindles are promising objective indicators in neurodegenerative disorders [2]. In order to interpret them, their structure needs to be clarified or a suitable model needs to be found. In this work, an autoregressive moving average (ARMA) model for sleep spindles is used to detect meaningful differences when applied to spindles from different types of people. More clearly, we wish to distinguish normal, elderly and dementia subjects based on the location of poles and zeros.

In [3] automated spindle detection by using autoregressive (AR) modelling for feature extraction is proposed. It is concluded that AR model parameters provide a good representation of the EEG data. It is expected that even better results can be obtained from the use of ARMA models.

In order to validate the ARMA models, a system is created and its response to white noise is obtained. Then, a model is estimated using the previous response and the estimated model is compared with the original one. It is also studied the best order of the model in order to represent a sleep spindle.

2 Contribution to Sustainability

Sustainability is to promote the best for people and environment, both now and in the future [8], and contributions to early diagnosis of diseases can lead to a better tomorrow. This paper comes with this perspective. The objective is an early detection of changes in brain to prevent or, at least, mitigate the influence of certain diseases.

3 Sleep Spindles

It is commonly referred in literature that sleep spindles are the most interesting hallmark of stage 2 sleep electroencephalograms (EEG) [1]. A sleep spindle is a burst of brain activity visible on an EEG and it consists of 11-15 Hz waves with duration between 0.5s and 2s in healthy adults, they are bilateral and synchronous in their appearance, with amplitude up to 30 μ V (Fig.1).

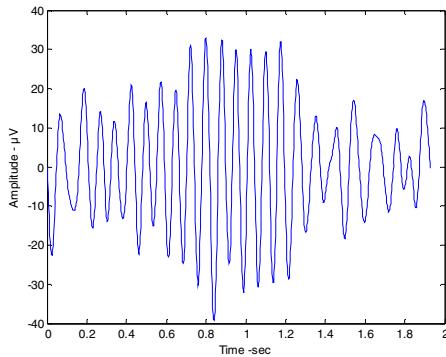


Fig. 1. EEG signal showing a sleep spindle

The spindle is characterized by progressively increasing, then gradually decreasing amplitude, which gives the waveform its characteristic name [4]. It is now reliable that sleep spindles are originated in the thalamus and can be recorded as potential changes at the cortical surface [5].

Sleep spindles are affected by brain pathology, as well as by normal and pathological aging (e.g., dementia) [1]. With normal aging, sleep spindles are less numerous and less well formed. In dementia, the sleep EEG patterns suggest accelerated aging [6].

Sleep EEG measures seem promising as objective indicators in neurodegenerative disorders, including dementia, where sleep changes appear to be an exaggeration of changes that come normally with aging.

4 ARMA Models and “Itakura-Saito” Distance

4.1 ARMA Model

In signal processing, autoregressive moving average (ARMA) models are typically applied to correlated time series data. Given a time series, we can consider it as the output of an ARMA system driven by white noise. The ARMA model is a tool for understanding and, whenever necessary, predicting future values in time series. The model consists of two parts, an autoregressive (AR) part and a moving average (MA) part. The model is usually referred to as ARMA(p,q) where p is the order of the autoregressive part and q is the order of the moving average part.

Compared with the pure MA or AR models, ARMA models more suitable for describing the characteristics of a given process with minimum number of parameters using both poles and zeros, rather than just poles or zeros [7].

As referred, a stationary ARMA process of order (p,q) is considered as the output of a linear time-invariant(LTI) digital filter driven by white noise. The transfer function of the system is given by:

$$H(z) = \frac{\sum_{m=0}^q b_m z^{-m}}{\sum_{k=0}^p a_k z^{-k}} \quad (1)$$

with $a_0=1$. The process corresponding to this model satisfies the difference equation:

$$x(n) = - \sum_{k=1}^p a_k x(n-k) + \sum_{m=0}^q b_m w(n-m) \quad (2)$$

where $w(n)$ is the input sequence, a zero-mean white noise and $x(n)$ is the output sequence. The main task in the modeling can be formulated as:

Given a segment of a time series, $x(n)$, $n=0,1,2 \dots, L-1$, estimate the $p+q+1$ ARMA parameters.

4.2 ARMA Model Validation

In order to validate the accuracy of the ARMA models some tests have been made.

Two models were tested, with 5 and 3 poles respectively and both with one zero. The models were excited with white Gaussian noise. An ARMA model was then estimated based only on the output of the model. The correct model orders are assumed to be known. The procedure was repeated several times (100) and means were calculated.

The original models used had the following transfer functions:

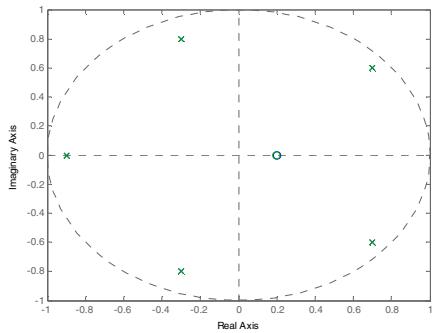
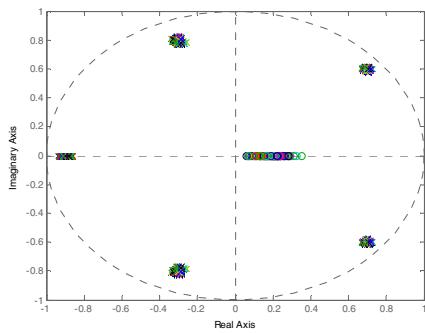
$$H1(z) = \frac{1 - 0.2 z^{-1}}{1 + 0.1 z^{-1} + 0.02 z^{-2} + 0.154 z^{-3} + 0.1597 z^{-4} + 0.5584 z^{-5}} \quad (3)$$

$$H2(z) = \frac{1 + 0.9 z^{-1}}{1 - z^{-1} + 0.66 z^{-2} - 0.4 z^{-3}} \quad (4)$$

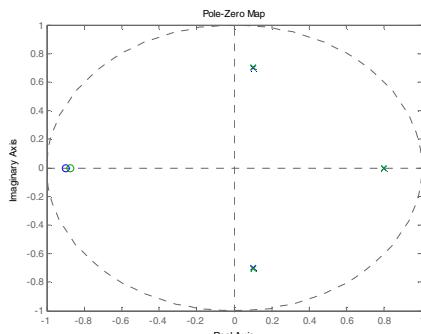
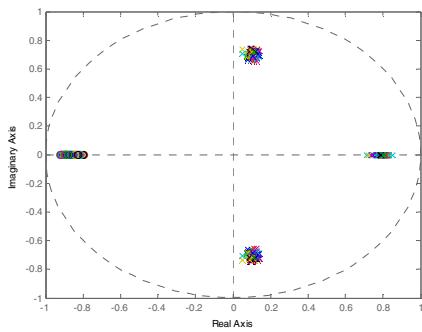
It can be seen, from (Tables 1 and 2) and from pole-zero map (Fig. 2 and 4) that the estimators produced very accurate results. In (Fig. 3 and 5) the clusters for poles and zeros positions are shown (these are the locations of all the poles and zeros in all the experiments).

Table 1. H1(z) - ARMA(5,1) coefficients

	a ₀	a ₁	b ₀	b ₁	b ₂	b ₃	b ₄	b ₅
Original	1.0000	-0.2000	1.0000	0.1000	0.0200	0.1540	0.1597	0.5584
Estimated(mean)	1.0000	-0.1973	1.0000	0.1027	0.0211	0.1514	0.1633	0.5602
Error	0.0000	0.0027	0.0000	0.0027	0.0011	0.0026	0.0036	0.0017
Mean of errors	0.0000	0.0481	0.0000	0.0402	0.0233	0.0198	0.0215	0.0221
Quadratic error	0.0000	0.0036	0.0000	0.0023	0.0009	0.0005	0.0007	0.0008

**Fig. 2.** Zeros and poles from original and estimated (mean) ARMA(5,1) systems**Fig. 3.** Clusters of zeros and poles from estimated ARMA(5,1) system**Table 2.** H2(z) - ARMA(5,1) coefficients

	a ₀	a ₁	b ₀	b ₁	b ₂	b ₃
Original	1.0000	0.9000	1.0000	-1.0000	0.6600	-0.4000
Estimated(mean)	1.0000	0.8736	1.0000	-1.0043	0.6737	-0.4045
Error	0.0000	0.0264	0.0000	0.0043	0.0137	0.0045
Mean of Errors	0.0000	0.0314	0.0000	0.0267	0.0345	0.0216
Quadratic error	0.0000	0.0017	0.0000	0.0011	0.0019	0.0007

**Fig. 4.** Zeros and poles from original and estimated (mean) ARMA(3,1) systems**Fig. 5.** Clusters of zeros and poles from estimated ARMA(3,1) system

In (Fig. 6) the Spectra from $H_2(z)$ and it's corresponding ARMA(3,1) model is show. It can be seen that both spectra are almost identical.

Tests have also been carried to determine the order of the model to be used in sleep spindle modelling. In (Figs. 4, 5, 7 and 8) the poles and zeros maps of 4 systems with different orders (numerator and denominator) are shown.

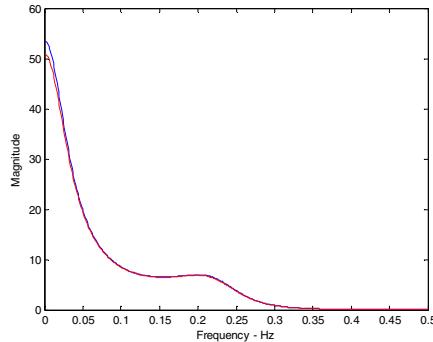


Fig. 6. Spectra from $H_2(z)$ and it's corresponding ARMA(3,1) model

In (Fig. 6) the Spectra from $H_2(z)$ and it's corresponding ARMA(3,1) model is show. It can be seen that both spectra are almost identical.

Tests have also been carried to determine the order of the model to be used in sleep spindle modelling. In (Figs. 4, 5, 7 and 8) the poles and zeros maps of 4 systems with different orders (numerator and denominator) are shown.

For systems with orders larger than 5 poles and 1 zero, the new poles or zeros tend to “accommodate” themselves to the system with minor differences in the overall model. For example, when one more zero is added, only the position of the other zero suffers notorious change to accommodate the new pole, with small variations in poles positions (Figs. 8 and 10). On the other hand, when we increase simultaneously the pole and zero orders, extra pole/zero pairs appear in very close positions or in reverse positions, revealing the presence of allpass subsystems.

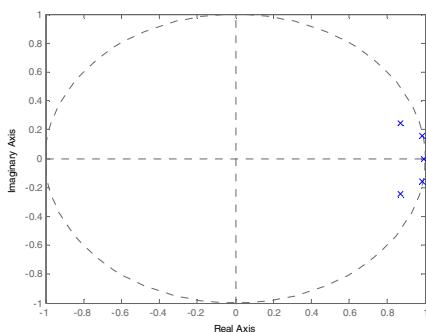


Fig. 7. Poles and zeros map of a spindle ARMA model with 5 poles, 0 zeros

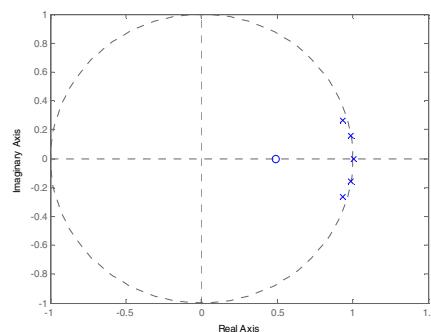


Fig. 8. Poles and zeros map of a spindle ARMA model with 5 poles, 1 zeros

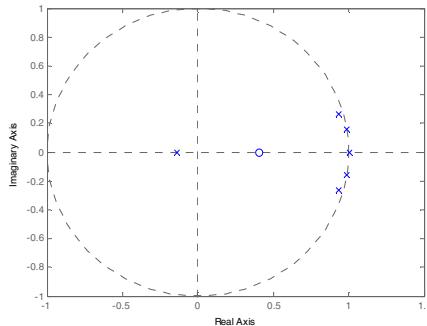


Fig. 9. Poles and zeros map of a spindle ARMA model with 6 poles, 1 zeros

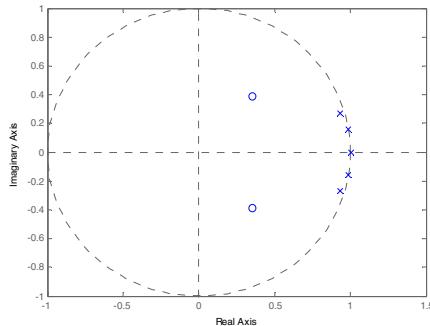


Fig. 10. Poles and zeros map of a spindle ARMA model with 5 poles, 2 zeros

4.3 “Itakura-Saito” Distances

The “Itakura-Saito” distance is a measure of the perceptual difference between original spectrum $P(W)$ and an approximation, $\hat{P}(w)$, of that spectrum. It can be used to compare the coefficients of the AR polynomials. It is defined as:

$$D_{IS}(P(w), \hat{P}(w)) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left[\frac{P(w)}{\hat{P}(w)} - \log \frac{P(w)}{\hat{P}(w)} - 1 \right] dw \quad (5)$$

5 Experimental Results

Spindles from night sleep of 5 subjects were used. Three sets of spindles from healthy subjects (S1, S2 and S3), a set of spindles from an elderly healthy subject (ELD) and a set of spindles from a dementia patient (DEM). The data used is from a real EEG with 512 Hz sampling rate. It has been pre-processed with a band-pass filter with cutoff frequencies of 5Hz and 22Hz.

For each person, the same procedure has been applied, consisting of:

- Visual identification of the sleep spindles;
- Estimation of an ARMA model with 5 poles and 1 zero, thus, obtaining A and B polynomials; the mean of A and B polynomials obtained from each set of spindles was computed;
- Zeros and Poles map of all systems were obtained;
- For computing the “Itakura-Saito” distances, the real poles and zeros were removed

It is possible to distinguish, either by the analysis of poles map (Fig. 11) or by “Itakura-Saito” distances healthy subjects from dementia/elderly subjects. It is particularly notorious the “zero” position, which in the elderly/dementia subjects is very close to -1. On the other side, normal subjects “zero” is found to be located on the right hand side complex plane (Fig. 8).

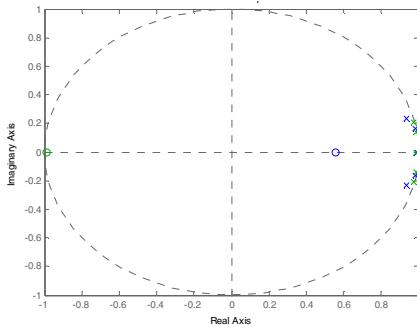


Fig. 11. Zeros and poles from elderly and normal subjects

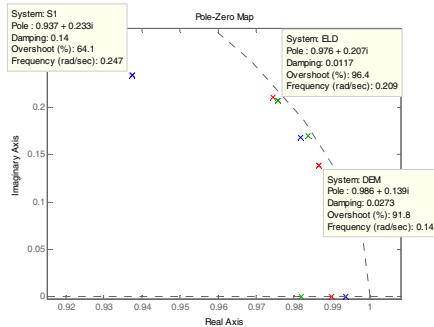


Fig. 12. Poles from elderly, dementia and normal subjects

However, it is not possible to distinguish between elderly and dementia subjects as the poles and zeros in both cases are very close to each others, as it can be seen from (Fig. 12).

The same result can be obtained by “Itakura-Saito” distance, in (Table 3) the distances between various subjects is showed. It is clearly seen that bigger distances (distance >0.1) are measured between elderly/dementia and normal subjects.

Table 3. “Itakura-Saito” distances between complex poles of subjects coefficients

	ELD	S1	S2	S3
DEM	0.0109	0.2235	0.1420	0.1044
S3	0.1037	0.0365	0.0134	
S2	0.1580	0.0777		
S1	0.1992			

It can be seen (Fig. 11 and 12) that the pole position give some biomarker for the presence of dementia. Complex poles from normal subject lie in specific areas, different from complex poles from elderly and dementia patients. However, pole location from elderly and dementia patients lie in similar regions inside the unit circle.

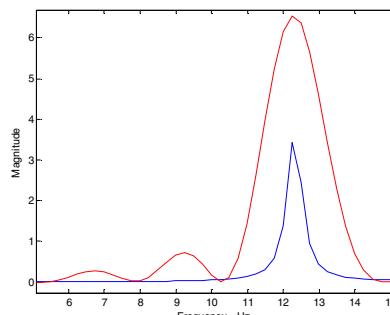


Fig. 13. Spectra from the ARMA model and from the original signal

In (Fig. 13) signal spectra corresponding ARMA model and to the periodogram estimate are shown. As it can be seen they give us similar information and from them we can conclude that the center frequency is 12.5 Hz, similar to a sinusoid in the 11-14 Hz band.

6 Conclusions

ARMA models can make a good representation of sleep spindles. It is showed that it is possible to distinguish between regular subjects and elderly or dementia subjects. However, it is not easy, using this method to distinguish between elderly and dementia subjects. According to [9,10] there is a increased loss of spindles in dementia patients when compared to elderly healthy subjects. From the experiments we performed it seems not to be possible to distinguish different abnormalities in the brain, probably because the effect on each individual spindle is similar. This requires further research.

This type of spindle modeling opens a door into the perspective of using it in the automatic spindle detection.

References

1. De Gennaro, L., Ferrara, M.: Sleep spindles: an overview. *Sleep Med. Rev.* 7, 423–440 (2003)
2. Ktonas, P.Y., Golemati, S., Xanthopoulos, P., Sakkalis, V., Ortigueira, M.D., et al.: Time-frequency analysis methods to quantify the time-varying microstructure of sleep EEG spindles: Possibility for dementia biomarkers? *J. of Neuroscience Methods* 185(1), 133–142 (2009)
3. Gorur, D., Halici, U., Aydin, H., Ongun, G., Ozgen, F., Leblebicioglu, K.: Sleep Spindles Detection Using Autoregressive Modelling. In: Kaynak, O., Alpaydin, E., Oja, E., Xu, L. (eds.) ICANN 2003 and ICONIP 2003. LNCS, vol. 2714. Springer, Heidelberg (2003)
4. Rechtschaffen, A., Kales, A.: A manual of standardised terminology, techniques and scoring system for sleep stages of human subjects. Public Health Service, Washington (1968)
5. Steriade, M., Jones, E.G., Llinas: Thalamic Oscillations and Signaling. Neuroscience Institute Publications. John Wiley & Sons, New York (1990)
6. Petit, D., Gagnon, J.F., Fantini, M.L., Ferini-Strambi, L., Montplaisir, J.: Sleep and quantitative EEG in neurodegenerative disorders. *J. Psychosom. Res.* 56, 487–496 (2004)
7. Kizilkaya, A., Kayran, A.H.: ARMA model parameter estimation based on the equivalent MA approach. *Digital Signal Processing* 16(6) (2006)
8. Wikipedia, <http://pt.wikipedia.org/wiki/Sustentabilidade>
9. Spinosa, M.J., Garzon, E.: Sleep spindles: validated concepts and breakthroughs. *J. Epilepsy Clin. Neurophysiol.* 13(4) (2007)
10. Reynolds III, C.F., Kupfer, D.J., Taska, L.S., et al.: EEG sleep in elderly depressed, demented, and healthy subjects. *Biological Psychiatry* 20(4), 431–442 (1985)

Pattern Recognition of the Household Water Consumption through Signal Analysis

Giovana Almeida¹, José Vieira², José Marques³, and Alberto Cardoso⁴

¹ University of Coimbra, PhD Stud., Dep. of Civil Engineering, Portugal
giovanaalmeida@dec.uc.pt

² University of Minho, Professor, Department of Civil Engineering, Portugal
jvieira@civil.uminho.pt

³ University of Coimbra, Professor, Department of Civil Engineering, Portugal
jasm@dec.uc.pt

⁴ University of Coimbra, Professor, Dep. of Informatics Engineering, Portugal
alberto@dei.uc.pt

Abstract. This paper presents the initial results of a research project that aims to develop a method for losses/leakage detection and household water consumption characterization through the detailed patterns analysis of signals generated by water meters. The Department of Civil Engineering (University of Coimbra) supports the research as part of a PhD Project. An experimental facility is used for signals acquisition and data analysis will be performed by using a pattern recognition algorithm that will identify the hydraulic devices in use. It is intended to develop and test some algorithm structures at various plumbing configuration forms to find the best one. In a second phase, a consumption analysis will be carried out using that algorithm to test its efficiency in inhabited houses. The expectation is to develop an efficient water monitoring tool that helps the users to follow-up and to control the water consumption using a computer or even a mobile device.

Keywords: pattern recognition, signal processing, monitoring household water consumption, efficient water use.

1 Introduction

A household telemetry system provides a more detailed and reliable water consumption data. For this reason it is being increasingly used by water supply companies as a management tool. Telemetry systems have several advantages and due to technological innovations the deploying cost is decreasing. However, the large number of data produced by daily measurements becomes a big challenge [1]. This creates opportunities for the development of models and algorithms using that information to give a detailed analysis of consumption and enabling an efficient water management. In order to contribute with efficient tools for the water consumption monitoring and control, the main goal of this research project from the Department of Civil Engineering (University of Coimbra) is to develop a method for losses/ leakage detection and water consumption characterization through the detailed study of the signals pattern generated by water meters.

This paper presents the developments of this research, focusing on the methodology, the state of the art and the work done so far. Till now a preparatory work on the controlled environment was carried out, including the calibration of the hydraulic system and the software development for signal acquisition, analysis and processing. After the laboratory stage we intend to analyze signals from data loggers installed in households to test and validate the new algorithms.

2 Contribution of Technological Innovation to Sustainability

During last century it has been observed that the demand of drinking water is growing faster (seven times) than the world population (four times) resulting in water scarcity risk in a several countries [2]. In Portugal, as well as in Mediterranean countries, it is estimated that in 10 or 15 years there will be lack of water quality due to the increase demand, inefficient water use and the wastewater mismanagement [3]. Annually the water losses in Portugal represent three billion of cubic meters of water and one half of this volume is lost in the urban environment, buildings and public systems of water supply [4]. The new technological tools that could contribute to the water losses control and water efficient uses are now essential to promote the water resources sustainability.

The main contribution of this research is to propose an algorithm to analyze in real time the online signal from a single water meter at the entrance of a house that enables to identify accurately the consumption pattern of each used device. As an online monitoring tool it is expected that it could be able to detect anomalies in consumption such as losses and excessive consumption of water in a timely and reliable manner. A computer or even a mobile phone can be used to access these data and contribute to the efficient water use and conservation of water resources.

3 Water Consumption Characterization (State of the Art)

The importance of detailed studies about water consumption characterization is growing due to the possibility to improve water management and its efficient use. The literature presents different methodologies for the consumption characterization. The accuracy of results varies according to each methodology, which is also reflected in the possibilities of analysis.

Researchers from USA [5], Spain [6] and Brazil [7] present different methodologies to characterize the water consumption in households. In these three research works were used water meters with pulsed output and dataloggers for data acquisition, however, different methodologies for data analysis were developed. The TraceWizard[©] 4.0 software was used in [5] for the specific recognition of signals through some parameters adjustment. To allow the signals identification it is necessary to know previously a set of properties for each device (flow, volume, duration of use, and others). This enables the program to distinguish between a tap use event and a toilet flush use event and so on. If these parameters are not well-adjusted, it is not possible to do the correct identification. When three or more events occur simultaneously it may not be possible to accurately disaggregate all end uses.

In [7], the pulses were converted into flow rate (l/s) to be possible to plot consumption graphs (flow vs. time). These graphs (representation of the water consumption signals) were correlated with time using the information of each device (supplied by

users). This correlation enabled to identify accurately the signals of some uses, but, due to the low accuracy of the provided information, some of the intakes had to be estimated. In [6], the information provided by the signal pulses was compared with the previous characterization of its amplitude and temporal patterns, which enabled to identify the water use in each moment. The low level of detail provided by that research report did not allow to estimate the methodology accuracy.

Studies [8] and [9] used a water meter on each hydraulic device and in the inlet pipe supply to guarantee good accuracy in the characterization phase. Despite the accuracy of the information, such methodology would be difficult to apply in inhabited buildings due to the work needed on the pipe supply installations. This situation does not have technical and economic viability in most of the situations.

4 Methods

The main goal of this study is to develop algorithms that make the pattern recognition of the water consumption behavior in each device using the data from the water meters at the inlet pipe. Experimental tests with controlled conditions were made to explore the several possibilities for setting up the facilities and to test algorithms. The steps of the experimental tests were: i) to build an experimental facility, ii) to calibrate the hydraulic system iii) to acquire the signals, iv) to store the signals, v) to process the data, and vi) to develop the algorithms for the water consumption pattern recognition.

Digital signal processing tools such as deconvolution algorithms [10] are used to data processing, aiming to solve the overlap of the signals coming from the simultaneous use of several devices in the hydraulic system. After that, techniques for feature extraction and classifiers to identify the signals of each flow classes in their respective equipment will be implemented and tested.

It is proposed to apply a feature extractor and a signal pattern recognition classifier to develop the algorithm according to the equipment studied. The features can be extracted in time or frequency domain. It is expected that the temporal or frequency characteristics of the transient response of the hydraulic system to the activation of a hydraulic devices vary according to their characteristics and their positions in the hydraulic supply system. This implies the need of a distinct signature requirement for each device even if we only have the flow rate signal in the pipe supply as an input for the classifier. It is expected that the algorithm compares the signal with the prototype generated for each device until the identification process is complete.

After the conclusion of the research in a controlled environment it is intended to carry out the data consumption analysis from inhabited houses to test and validate the algorithm. The data signals from the intake will be analyzed for one year, covering the four seasons.

5 Preparation of the Experimental Apparatus

The considered hydraulic system includes two volumetric water meters with pulsed output (Actaris Aquadis +, class D) and two taps. The counted volume is converted in a pulse sequence through a sensor (CybleTM Sensor) installed on the water meter. A data acquisition card with USB interface (USB card NIidaq 6009 data acquisition - DAQ)

interconnect the Cyble™ Sensor and the laptop. This card allows the data acquisition through various analog input channels and stores the data using MatLab/Simulink® software. Some features were developed in Simulink® for data acquisition and processing enabling the analysis and development of classification algorithms based on the flow rate signal.

In future developments a pressure transducer will be used in this experimental facility to check how changes in pressure within the pipe can affect the signals characteristics. Furthermore, the localized head losses will also be studied since, theoretically, these losses also interfere with these signals.

6 Tools for Data Acquisition and Signals Analysis

The low-frequencies (LF) signal transmitted by the Cyble™ Sensor detects each rotation¹ of the water meter and then it emits 1 pulse per revolution. It remains active whenever there is a flow, whatever the flow direction is. Each sensor is connect to an analog input (AI) channels of the DAQ card using a two wired cable. After the analog-to-digital conversion (ADC) in the DAQ card, the data in digital format is send to the computer through a USB cable.

To make sure that a discrete signal is representative of an analog signal, it is required to verify the Nyquist–Shannon sampling theorem [11] which says that the sampling rate must be greater than twice the highest frequency of the signal. In this case, to allow the identification of each water meter transition, whenever it counts one volume, having been considered in the experiments a sampling frequency of 50 samples/sec.

The functionalities for data collecting and processing at the computer were developed in MatLab/Simulink® which is a commercial tool for modeling, simulation and analysis of dynamic systems. Simulink® tool uses a graphical representation of blocks as the main interface, as well as a customizable set of block libraries. It offers integration with the rest of MatLab® and it is widely used in control theory and digital signal processing for simulation of systems in various domains [12].

The data acquisition block with multiple analog input channels was used as interface in the Simulink® model. The water meter signals were connected to two input channels and they are represented by the signals HWChannel0 and HWChannel1, as shown in Fig. 1. Output blocks for data storage in the Matlab® workspace (pulse signals) and a Scope block (to show the signal during the data acquisition) were also considered. To distinguish the signals from the water meters, the analog output block generates two different values for the reference voltage of each pulse signal (3.5V and 2.5V for the signals from the water meter 1 and 2, respectively). A clock block was considered to generate the current simulation time. The Simulink® data acquisition block was configured selecting the channels involved (channels 0 and 1), the sampling frequency (50 samples/s) and the number of samples provided by the DAQ for each channel (at least 2 samples per channel).

¹ In each rotation pass through the water meter 0.1 litre of water.

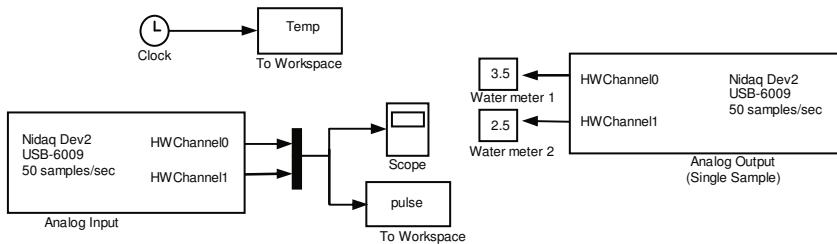


Fig. 1. Simulink® Data Acquisition Model

Given that the DAQ card sends pairs of samples to the computer, it was necessary to develop a MatLab® function to arrange data sequentially according to the sampling time. Until this point the signals are acquired as data pulses. However, to enable the signal analysis using pattern recognition it is necessary to convert them to discrete time flow rate signals. Therefore, another Simulink® model was developed to convert data pulse in data flow rate (Fig. 2).

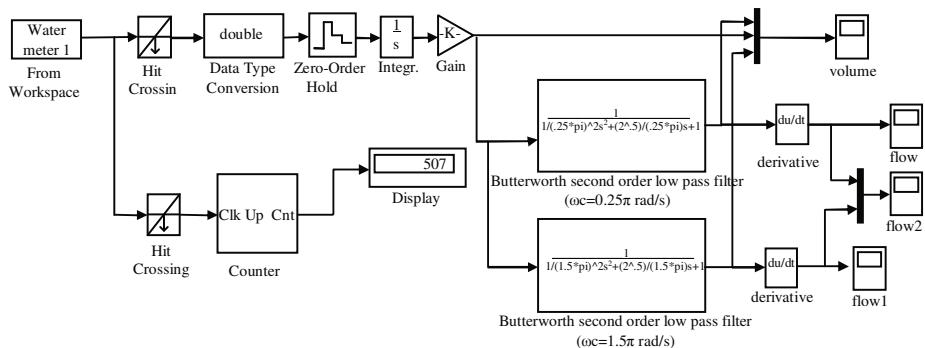


Fig. 2. Simulink® volume and flow rate calculation model

After reading the pulses from the MatLab® workspace, they are counted through the “hit crossing” block. The “hit crossing” block detects when the input signal reaches the parameter value (0.1) in the direction specified (falling) and output 1 in the crossing time. The “counter” block counts how many times the parameter value is reached and presents it in display. The “data type conversion” block converts the input data to a suitable data type for integration and the “zero order hold” block converts the discrete-time signal to a continuous-time signal. The volume is calculated through the “integration” of the data pulse considering a 0.1 liters and 50 samples/sec as a “gain”. The volume (Fig.3), flow (Fig.5), flow1 (Fig.4) and flow2 blocks plot the signals resulting from the simulation process.

A low-pass filter was used to cut unwanted frequencies and to smooth the graph of volume (stairs shape) allowing the calculation of its derivative to get the flow rate curve. The stair shape of the volume graph arises due to the sequence of discrete values generated by the rotating piston of the volumetric water meter (fig.3). A second order

Butterworth low-pass filter with cutoff angular frequencies (ω_c) of 1.5π rad/s and 0.25π rad/s were used. The transfer function $H(s)$ used in this filter is given by:

$$H(s) = \frac{1}{\left(\left(\frac{1}{\omega_c}\right)^2 s^2 + 1,4142\left(\frac{1}{\omega_c}\right)s + 1\right)} \quad (1)$$

After the application of the filter, smoother curves were obtained as seen in Fig. 3. With the derivative calculation it was possible to obtain the corresponding flow rate signal.

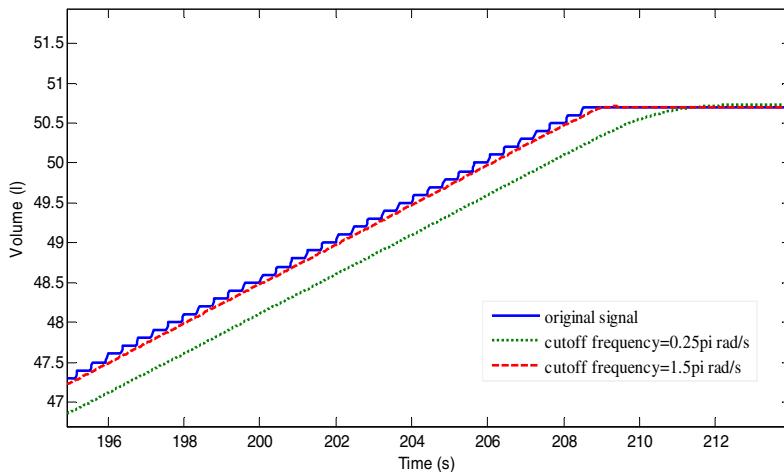


Fig. 3. Volume graph with application of low-pass filters

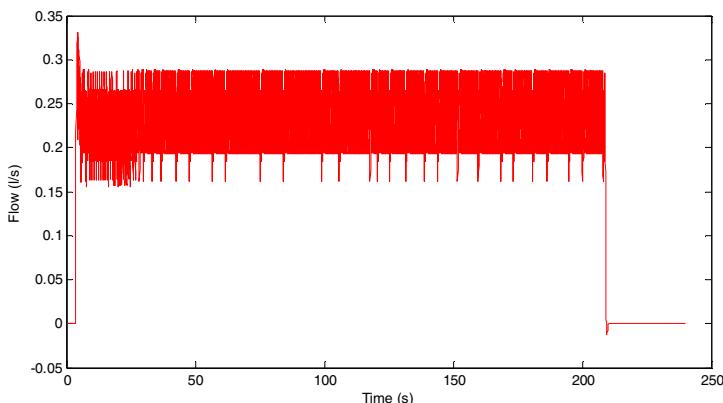


Fig. 4. Flow rate graph using a low-pass filter with 1.5π rad/s cutoff angular frequency

Although the values of the resulted chart using the low-pass filter with 1.5π rad/s cutoff angular frequency are close to the original values of the volume graph, it still keeps some frequency components that interfere with the corresponding flow rate signal (Fig. 4). The filter with 0.25π rad/s cutoff angular frequency provides smoother curves highlighting their characteristics (Fig. 5).

Comparing the volumes recorded in the water meters with those obtained by the models developed in MatLab/Simulink® it was observed that the values are similar. Therefore, it can be concluded that the developed models are adequate to the goals.

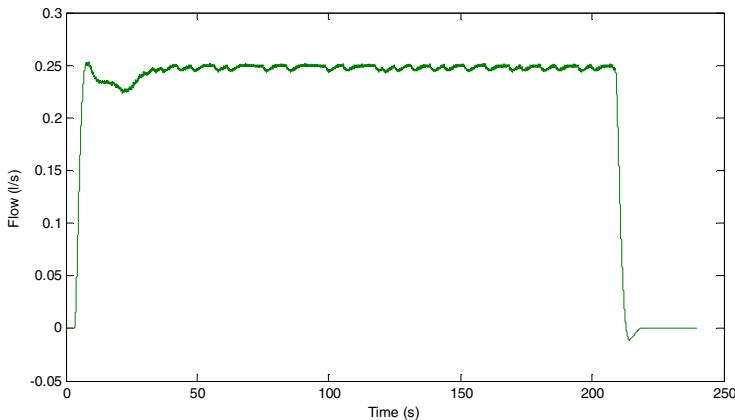


Fig. 5. Flow rate graph using a low-pass filter with 0.25π rad/s cutoff angular frequency

7 Final Considerations

Until now the work was mainly focused on the experimental facility preparation, on the hydraulic system calibration and on the software development for signal acquisition, analysis and processing. The results obtained so far are considered satisfactory, since the equipment installed and the tools developed are in full operation, meeting the objectives. The next goal is to analyze a variety of hydraulic flow rate signals in different configurations considering a feature extractor. Based on this analysis, pattern recognition algorithms will be developed, where those with the highest number of hits will be chosen. After the laboratory stage it is intended to use dataloggers installed in inhabited houses to acquire signals to test and validate the algorithm in a real environment.

It is expected that the results of this research will be considered as a relevant improvement for an efficient control of water consumption using a computer or even a mobile device that can be followed by users in any period of time (hourly, daily, weekly, ...). It is also expected that unnecessary expenses, resulting from water losses, like leaks or ruptures, or water overuse, will be identified, developing, for instance, a personal water consumption monitoring system or even an automatic system to generate alarms whenever an abnormal consumption pattern is detected.

Acknowledgments. This work is being funded by the Foundation for Science and Technology (FCT) as a PhD grant.

References

1. Medeiros, N., Loureiro D., Mugeiro J., Coelho S.T., Branco, S.: Concepção, Instalação e exploração de sistemas de telemetria domiciliária para apoio à gestão técnica de sistemas de distribuição de água. In: I Conferência INSSAA – Modelação de Sistemas de Abastecimento de Água, Barcelos (2007)
2. Human Development Report 2006. UNDP,
<http://hdr.undp.org/en/reports/global/hdr2006>
3. Observatório do Algarve,
http://www.observatoriodoalgarve.com/cna/noticias_ver.asp?noticias=16003
4. Diário Digital,
http://diariodigital.sapo.pt/news.asp?section_id=114&id_news=459108&page=0
5. Mayer, P.: Water And Energy Savings From High Efficiency Fixtures and Appliances in Single Family Homes. USEPA — Combined Retrofit Report 1 (2005)
6. González, F.C., Rueda, T.R., e Les, S.O.: Microcomponentes y factores explicativos del consumo doméstico de agua en la Comunidad de Madrid. Cuadernos I+D+I 4. Canal de Isabel II (2008)
7. Almeida, G.A., Kiperstok, A., Dias, M., Ludwig, O.: Metodologia para Caracterização de Consumo de Água Doméstico por Equipamento Hidráulico. Anais do Silubesa/ Abes. Figueira da Foz (2006)
8. Barreto, D.: Perfil do consumo residencial e usos finais da água. Ambiente Construído, Porto Alegre 8(2), 23–40 (2008) ISSN 1678-8621; © 2008, Associação Nacional de Tecnologia do Ambiente Construído (April/June 2008)
9. Fernandes, B.C.: Construção de um Sistema Eletrônico de Monitoramento de Consumo de Água Residencial. Projeto de Graduação apresentado ao Departamento de Engenharia Elétrica. p. 65 Centro Tecnológico da Univ. Federal do Espírito Santo (2007)
10. Diniz, P.S.R., Silva, E.A.B., Netto, S.L.: Processamento Digital de Sinais: Projeto e Análise de Sistemas. Bookman, Porto Alegre-RS (2004)
11. Ifeachor, E.C., Jervis, B.W.: Digital signal processing: a practical approach, p. 760. Addison-Wesley, Wokingham (1993)
12. Mathworks,
<http://www.mathworks.com/products/simulink/description1.html>

Survey on Fault-Tolerant Diagnosis and Control Systems Applied to Multi-motor Electric Vehicles

Alexandre Silveira¹, Rui Esteves Araújo², and Ricardo de Castro²

¹ Instituto Superior de Engenharia do Porto

² Faculdade de Engenharia, Universidade do Porto, Portugal

asi@isep.ipp.pt, {raraajo, de.castro}@fe.up.pt

Abstract. In the last years we have witnessed a growing interest, by the academic community and the automotive industry, in the multi-motor electric vehicles. The electrical nature of the propulsion is going to stress even more an increasing insertion of electronic devices in the vehicles. Furthermore, carmakers are performing research and already presented some vehicles based on the concept of X-By-Wire. Consequently, the growing complexity of the actuators and their control, as well as the need of increasing the safety and reliability of the vehicles obliges to the study and development of intelligent computational systems dedicated to the detection and diagnosis of failures in the electric propulsion. Hence, it is fundamental to start advanced studies leading to the development of innovative solutions that embed fault-tolerant electric propulsion in the electric vehicles. Accordingly, the main objective of this work consists on the bibliographic revision and study of fault-tolerant diagnosis and control systems dedicated to multi-motor electric vehicles.

Keywords: Fault detection and diagnosis, Fault tolerant control systems, multi-motor electric vehicles.

1 Introduction

As electric vehicles are systems with propellers based in electromechanical drives, we can classify them as critical systems, where the use of fault-tolerant control techniques becomes essential. As referred in [1] and [2], the interest of introducing fault detection, tolerance and redundancy in a system is to increase its safety and reliability. A system is considered to be safe if it is able to prevent any danger to humans, equipments and environment; and is reliable if it is capable of perform correctly the required functions, over a certain period of time under a given set of conditions. These characteristics are of great importance in safety related processes and systems like aircrafts, trains, automobiles and power plants [2].

So, in order to improve system's reliability and operational safety, reducing the possibility of those failures or trying to predict its happening before occurrence it's necessary. One way to do this is to employ Fault Detection and Diagnosis systems (FDD). The FDD consists of making a binary decision when something wrong happens, and to determine the location and the nature of the fault. These methods are

based on the concept of redundancy, which can be obtained with hardware or using analytical redundancy. In the former, backup sensors and/or actuators are used in such a manner that the system is able to automatically replace the faulty ones when a given fault is identified. This strategy is not always possible due to physical or economical constrains. An alternative way to turn the system more reliable is to use the concept of analytical redundancy that uses a mathematical model and estimation techniques. With this kind of concept, the system is able to shut down itself or to employ adequate procedures to tolerate the faults and maintain the system operational [3].

This paper is structured as follows: In section 2 it is shown the contribution to sustainability of this paper. In section 3 it is made an overview on the Fault Detection and Diagnosis approaches for implementation in electric vehicles (EVs) and in section 4 there are presented some conclusions about this survey.

2 Contribution to Sustainability

The interest in the electric vehicles rose recently due both to environmental questions and to energetic dependence of the contemporary society [4]. Moreover, in the passenger car industry, the majority of the constructors are currently developing considerable efforts to introduce the first generation of pure electric vehicles, like the *Nissan Leaf*, the *Mitsubishi iMIEV* or the *Fluence Z.E.* of Renault. In fact, some are already available in some European markets. Accordingly, it is necessary to study and implement in these new vehicles fault-tolerant control systems which enable them to be more reliable and safe, enhancing its sustainability.

3 Overview on Fault Detection and Diagnosis for EVs

The fault detection and diagnosis techniques are well studied in the specialty literature [1], [5] and [6], and are more and more applied in industrial processes and products. Therefore, the actual state of the technique motivates the next step, which consists in the following: after the fault detection the controller must be capable of guarantee the functioning of the vehicle in safety conditions. In other words, it is necessary to study and develop fault-tolerant control architectures dedicated for the future electric vehicles.

Lately, as we are getting more and more concerned about safety, reliability and sustainability, there was a rise in the research of fault detection and diagnosis systems, that led to the development of many FDD techniques [1], [7]. Although these technical and scientific progresses, the conception of fault-tolerant controllers oriented to electric vehicles is simultaneously a complex and fascinating project. The reasons for this are inherent to the difficulties of controlling in an efficient, effective and secure way the several propellants in a multi-motor electric vehicle.

Detailed analysis of faults processes has indicated the need to act quickly following a device failure to prevent propagation of faults that may lead to catastrophic failure of the propulsion system. To minimize the effect of fault, it is essential to accurately identify the failed devices and its mode of failure. Historically, in what concerns practical applications, a great amount of research on fault-tolerant control systems

was derived from the aerospace industry [1]. Nowadays is recognized some maturity to the theoretical concepts of the fault-tolerant controllers, whose one of the pioneers was Ron Patton [8], where he makes a bibliographical revision encompassing the principal fault-tolerant systems control areas.

The works of Mutoh and Nakano [9], demonstrate by simulations that for electric multi-motor vehicles, faults in one of the propulsion systems lead to the loss of stability of the vehicle. Thus, their work motivates for the necessity of study and evaluation of different configurations regarding the number of motors and their localization in the vehicle. The possibility of the inclusion of the motor in the wheel will contribute to the conceptualization, validation and implementation of new innovative active torque distribution methodologies for the future vehicles, potentiating improvements in their handling, safety and stability. However, this innovative solutions lead to new challenges due to the fact that the electric motors are feed by electronic power converters that have a bigger fault probability. In the case of Electric Hybrid Vehicles, if the motor is not monitored, motor faults might lead to severe damages or even accidents [10].

Considering the electric vehicles as critical systems, with propellers based in electromechanical drives, the use of fault-tolerant control techniques is essential. On the other hand, the adoption of the “X-by-wire” technology, where “X” represents the several subsystems to control in the vehicle, as “Steer-by-wire”, “Throttle-by-wire” or “Brake-by-wire”, motivate to the study of fault-tolerant controllers [11], [12].

In the literature we can find several solutions in the thematic of the fault-tolerant electromechanical actuators, with special emphasis in the electronic converters and motors [13]. The work of Delgado [14] presents a review of fault-tolerant systems of electronic speed drives to DC motors and induction motors and some hardware configurations. Also, Murphrey et al. [15] presents a fault diagnosis Neural Network system that detects and locates multiple classes of faults in electric drives, using a machine learning technology.

Regarding traction applications, Akin [10] presents a fault diagnosis system to electric or hybrid vehicle's motors, based on the Fast-Fourier Transform (FFT) method. In four wheel steer/four wheel drive systems (4WS4WD) we must stress the work of Yang [16] that proposes a fault-tolerant hybrid approach to maintain the vehicle in a functional state even in the presence of a failure. Thus, there is an increasing demand for dynamic systems to operate autonomously in the presence of faults and failures in sensors, actuators and components. So, fault detection and diagnosis are essential components in an autonomous fault tolerant control system [17]. Therefore it is necessary to design control systems capable of tolerate possible faults in those systems, in order to improve reliability and availability [18].

A Fault-Tolerant Control Systems (FTCS) is a control system capable of accommodating system component faults (actuators, sensors) and able to guarantee stability and an acceptable degree of performance. FTCS can also prevent that faults in a subsystem may develop into failures at the system level [18]. Fault-Tolerant Control Systems (FTCS) can be classified into two types: passive (PFTCS) and active (AFTCS). Contrasting with the former, the latter react with system component failures actively and implement reconfiguring actions to maintain the system stable and with acceptable performance [1]. AFTCS are usually constituted of four subsystems: a reconfigurable controller, a FDD scheme, a controller reconfiguration mechanism and a command/reference governor [1], as illustrated in Fig. 1.

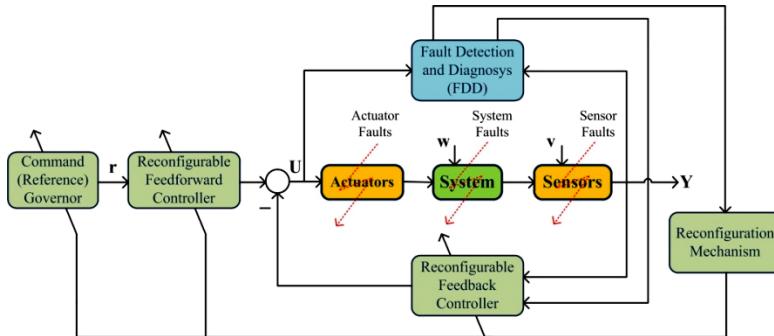


Fig. 1. General structure of AFTCS [1]

The FDD block represented above is capable of detecting the presence of faults in the system that it monitors, is able to determine their locations and can estimate their severities. In other words, it is capable of doing three tasks [17]:

- **Fault detection:** to decide if everything is working like expected or something has gone wrong;
- **Fault isolation:** to determine the location (component, sensor, actuator) of the fault;
- **Fault identification:** to estimate the severity, type or nature of the fault.

The reconfigurable controller should be designed automatically to compensate the fault-induced changes in the system in order to maintain their stability and closed-loop system performance [1].

In the recent years, many research and work have been done in the area of FDD. One example is the paper of Zhang and Jiang [1] where they presented a classification of several FDD methods. As explained by them, FDD approaches are usually classified into two categories: (1) model-based and (2) data-based (model-free) schemes. Moreover, these two schemes can also be classified into qualitative and quantitative approaches. Relatively to this classification, Muenchhof [7] and Liu [19] refer that classical model-free methods rely on hardware redundancy, resulting in extra hardware, cost and size but on the other hand can result in reduction of unexpected downtime of the system. Contrarily, model-based methods rely on analytical redundancy, where consistency between the expected behavior and measurements of the process is checked based on analytical models [19]. As reported in [18], model-based methods are best suited for processes which input and output signals can be measured. However, if we can only measure the outputs of a process, signal-based methods should be applied.

3.1 Model-Based Methods

Model-based fault detection methods are well studied and reported in the literature [2], [7], [19], [20], [21] and [22]. According to [18], model-based methods are widely used and are usually performed in two steps: residual generation and residual evaluation. In this kind of method, the difference between the measurement and the

expected behavior is called residual. As mentioned by Isermann [2], this consists of the detection of faults of processes, actuators and sensors by using dependences, expressed by mathematical process models, between different measurable signals. The basic structure of model-based fault detection is illustrated in Fig. 2.

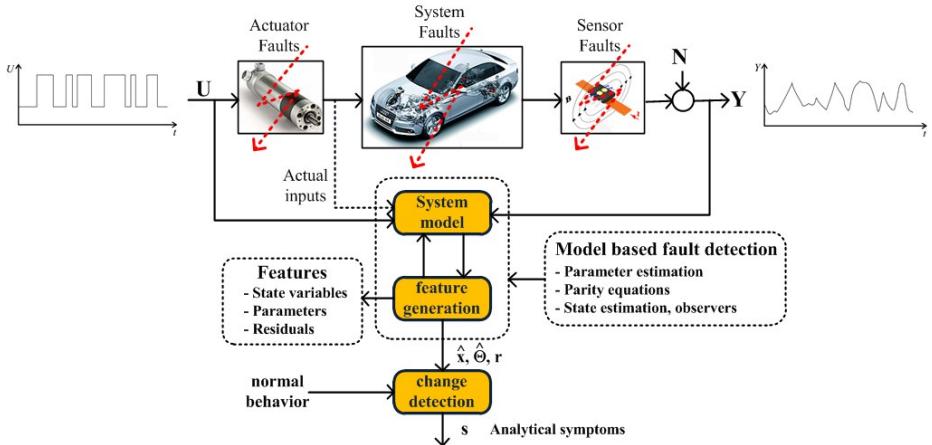


Fig. 2. Basic structure of model-based and signal based FD [2]

Based on the measured input signals U and output signals Y , the detection methods generate the residuals r , parameter estimates $\hat{\Theta}$ or state estimates \hat{x} , which are called features. Comparing them with normal features, changes are detected, leading to analytical symptoms s [2]. This analytical symptom generation is quantifiable and analytical information about the process, and is a result from the data processing of the measured signals [18]. We can divide model-based methods in four main classes: state estimation approaches, parity space approaches, parameter estimation approaches and simultaneous state/parameter estimation approaches. In first class, the system outputs are estimated from measurements using: Luenberger observer, linear or nonlinear observers, sliding-mode observers and high-gain nonlinear observers, for deterministic cases. In the case of stochastic ones, outputs are estimated using: Kalman filter (linear, extended and unscented) or receding horizon estimators [17] and [18].

In the parity space approaches, residual are computed as difference between measured outputs and estimated outputs and their associated derivatives. In the parameter estimation approach, residuals are computed as the parameter estimation error, by continuously estimating the parameters of a process model [18]. Parity space methods are based on simple algebraic projections and geometry and are more sensitive to measurement noise and process noise as compared to observer-based methods [17]. Still, these methods are mainly suitable for detection and isolation of additive faults. They are simpler and easy to implement compared to observer based techniques. Parameter estimation approach is based on the assumption that the faults are reflected in the physical parameters of the system. In this approach the system parameters are estimated online with parameter estimation techniques [17].

In this model-based approach there are also some works related to the problematic of safety in X-by-Wire systems such as drive-by-wire, steer-by-wire, throttle-by-wire or brake-by-wire [11], [12], [21] [23], and [24].

Most of fault-tolerant control systems require hardware redundancy to make them more reliable [11], [12], [24] and [25]. This means that one or more modules are connected, usually in parallel. Such redundant schemes can be implemented for hardware, software, information processing and mechanical and electric components like sensors, actuators, microcomputers or power supplies [12]. Regarding these redundant structures, there are two basic approaches for fault tolerance: (1) static redundancy and (2) dynamic redundancy. Static Redundancy uses multiple redundant modules with majority voting (all modules are active). Dynamic redundancy requires fewer modules at the cost of more information processing. However, without mechanical redundancy, the reliability of the system needs to be improved by implementing fault-tolerant control techniques [21]. In fact, with the profit margin already low, mechanical redundancy will not be acceptable to the automobile industries.

Employing analytical redundancy techniques instead of hardware redundancy it will be possible to reduce costs, volume and weight to a point where the automakers feel comfortable, without compromising safety and reliability required by consumers [26]. According to [20] the concept analytical redundancy stands generally for an analytical reconstruction of quantities or parts of the system or process under monitoring. For Anwar [24], the concept of analytical redundancy has been investigated to replace hardware redundancy, in order to reduce overall cost and at the same time to improve reliability.

The above discussed techniques of model based approaches to fault diagnosis are powerful if an accurate system model is available. The diagnostic performance may be limited when it is not possible to obtain accurate and robust models.

3.2 Data-Based Methods

An alternative approach to the model-based residual generation is the data-based approach, also called model-free approach. This learning-based method learns the plant model from an historical input and output data of the system, i.e., it detects faults by analyzing specific properties of measured signals. According to [17] data-based methods are based on signal processing techniques, obtained using either of the following two types: (1) Time domain limit checking and/or trend analysis and (2) Frequency or mixed time-frequency domain analysis. In the former, the statistics of the measurable states and outputs of the system are compared with nominal operating limits. In the latter, it is made an analysis of the time series of system states and outputs measured by system sensors [17]. In the time domain, the most common technique is the Qualitative Trend Analysis (QTA), while in frequency domain the most widely used algorithms are the Discrete Fourier Transform (DFT), and the Discrete Wavelet Transform (DWT) [17].

For [17], the major drawback of these signal processing techniques is that they do not consider the dynamic interrelationship between the different measured signals of the system. Thus, these techniques are more appealing for situations where high-fidelity mathematical model of the monitored system does not exist or is very difficult

to obtain. According to the same author, artificial neural networks, fuzzy logic and neuro-fuzzy systems are the most widely used approaches. Using these techniques we can work on the quantitative and qualitative information of the monitored system. Qualitative information is expressed in the form of Boolean or fuzzy if-then rules. These have a drawback that is the problem of deriving Boolean or fuzzy if-then rules in many engineering applications. In fact, this requires extensive expert knowledge of the system [17]. On the other hand, Neural Networks (NN) are ideal mathematical tools for situations like this, where the knowledge that describes the behavior of the system is stored in large quantitative datasets [17]. Also, as mentioned by Patton [27], a well trained NN has the capacity of making intelligent decisions even in the presence of noise, system disturbances and corrupted data.

4 Conclusions

Fault Detection and Diagnosis are systems of great importance for modern electric vehicles. This is even more critical for multi-motor electric vehicles, since their stability is deteriorated in the presence of fault in one of the thrusters. As their thrusters are constituted by the electronic power converters and the motors, it is necessary to study the faults in these systems and to implement FTCS to deal with them, maintaining safety for their users and ensure sustainable operation.

In summary, even though there are some pioneering works, there are not known published ones describing the application of fault-tolerant control techniques, especially with a holistic view on the multi-motor electric vehicles.

References

1. Zhang, Y., Jiang, J.: Bibliographical review on reconfigurable fault tolerant control systems. *Annual Reviews in Control*, 229–252 (2008)
2. Isermann, R.: Model-based fault-detection and diagnosis - status and applications. *Annual Reviews in Control*, 71–85 (2004)
3. Li, X.: Fault Detection Filter Design for Linear Systems, PhD Dissertation (August 2009)
4. Chan, C., Chen, K.: Electric, Hybrid, and Fuel-Cell Vehicles: Architectures and Modeling. *IEEE Transactions on Vehicular Technology* 59, 589–598 (2010)
5. Blanke, M., Kinnaert, M., Lunze, J., Staroswiecki, M.: *Diagnosis and Fault Tolerant Control*, 2nd edn. Springer, Heidelberg (2006)
6. Patton, R., et al.: *Issues of Fault Diagnosis for Dynamic Systems*. Springer, Heidelberg (2000)
7. Muenchhof, M., Beck, M., Isermann, R.: Fault-tolerant actuators and drives - Structures, fault detection principles and applications. *Annual Reviews in Control* (2009)
8. Patton, R.: Fault-tolerant control: the 1997 situation. In: IFAC Symposium on Fault Detection, Supervision and Safety for Technical Processes (SAFEPROCESS1997), Hull, UK, vol. 2, pp. 1033–1055 (1997)
9. Mutoh, N., Nakano, Y.: Dynamic Characteristic Analysis of the Front- and Rear-Wheel Independent-Drive-Type Electrical Vehicle (FRID EV) When the Drive System Failed during Running under Various Road Conditions. In: IEEE Vehicle Power and Propulsion Conference VPPC 2009 (2009)

10. Akin, B., Ozturk, S.B., Toliyat, H.A., Rayner, M.: DSP-Based Sensorless Electric Motor Fault-Diagnosis Tools for Electric and Hybrid Electric Vehicle Powertrain Applications. *IEEE Transactions on Vehicular Technology* 58(6) (2009)
11. Naidu, M., Gopalakrishnan, S., Nehl, T.: Fault Tolerant Permanent Magnet Motor Drive Topologies for Automotive X-By-Wire Systems (2009)
12. Iserman, R., Schwarz, R., Stölzl, S.: Fault tolerant drive-by-wire systems. *IEEE Control Systems Magazine*, 64–81 (2002)
13. El Hachemi Benbouzid, M., Diallo, D., Zeraoulia, M.: Advanced Fault-Tolerance Control of Induction-Motor Drives for EV/HEV Traction Applications: From Conventional to Modern and Intelligent Control Techniques. *IEEE Transactions on Vehicular Technology* 56(2) (2007)
14. Delgado, D., et al.: Fault-Tolerance Control in Variable Speed Drives: A Survey. *IET Electric Power Applications* 2, 121–134 (2008)
15. Murphey, Y.L., Masrur, M.A., Chen, Z., Zhang, B.: Model-based fault diagnosis in electric drives using machine learning. *IEEE/ASME Transactions on Mechatronics* 11(3), 290–303 (2006)
16. Yang, H., Cocquempot, V., Jiang, B.: Hybrid Fault Tolerant Tracking Control Design for Electrical Vehicles. In: *IEEE 16th Mediterranean Conference on Control and Automation* (2008)
17. Sobhani-Tehrani, E., Khorasani, K.: Fault Diagnosis of Nonlinear Systems using Hybrid Approach. Springer, Heidelberg (2009)
18. Mahmoud, M.M., Jiang, J., Zhang, Y.: Active Fault Tolerant Control Systems: Stochastic Analysis and Synthesis. Springer, Heidelberg (2003)
19. Liu, L., Logan, K.P., Cartes, D.A., Srivastava, S.K.: Fault Detection, Diagnostics, and Prognostics: Software Agent Solutions. *IEEE Transactions on Vehicular Technology* 56(4), 1613–1622 (2007)
20. Ding, S.X.: Model-based Fault Diagnosis Techniques: Design Schemes, Algorithms, and Tools. Springer, Heidelberg (2008)
21. Im, J.S., Ozaki, F., Yue, T.-K., Kawaji, S.: Model-based fault detection and isolation in steer-by-wire vehicle using sliding mode observer. *Mechanical Science and Technology*, 1991–1999 (2009)
22. Dumont, P.E.: Fault Detection of Actuator Faults for Electric Vehicle. In: *16th IEEE International Conference on Control Applications*, Singapore, pp. 1067–1072 (2007)
23. Laboratory, Army Research. Using a Steering Shaping Function to Improve Human Performance in By-Wire Vehicles. Document (August 2009),
<http://www.defensetechbriefs.com/component/content/article/5554>
24. Anwar, S., Chen, L.: An Analytical Redundancy-Based Fault Detection and Isolation Algorithm for a Road-Wheel Control Subsystem in a Steer-By-Wire System. *IEEE Transactions on Vehicular Technology* 56(5), 2859–2869 (2007)
25. Hwang, I., Kim, S., Kim, Y., Chze, E.S.: A Survey of Fault Detection, Isolation, and Reconfiguration Methods. *IEEE Transactions on Control Systems Technology* 18(3), 636–652 (2010)
26. Hasan, M.S., Anwar, S.: Sliding Mode Observer Based Predictive Fault Diagnosis of a Steer-By-Wire System. In: *17th World Congress of the International Federation of Automatic Control*, Seoul, Korea, pp. 8534–8539 (2008)
27. Patton, R.J., Lopez-Toribio, C.J.: Artificial intelligence approaches to fault diagnosis. *IEEE Colloquium on Update on Developments in Intelligent Control* (Ref. No. 1998/513), 311–312 (October 23, 1998)

Design of Active Holonic Fault-Tolerant Control Systems

Robson M. da Silva¹, Paulo E. Miyagi², and Diolino J. Santos Filho²

¹ State University of Santa Cruz, Rod. Ilhéus/Itabuna, km 16, CEP 45662-900 Ilhéus, BA

² University of São Paulo, Av. Prof. Mello Moraes, 2231, CEP 05508-030 São Paulo, SP Brazil

rmsilva@uesc.br, pemiyagi@usp.br, diolino.santos@poli.usp.br

Abstract. The adequate evaluation of new technologies in productive systems that perform multiple and simultaneous processes, exploring the intense sharing of resources, demands the updating of supervision and control systems. On the other hand, totally infallible systems are unviable, and for a flexible productive system (FPS) does not suffer interruptions due to component failure, the concept of AFTCS (*active fault-tolerant control system*) mechanism must be adopted. In this sense, holonic control system (HCS) is considered a trend for an intelligent automation and the combination of HCS techniques and AFTCS is fundamental to assure efficiency, flexibility and robustness of FPSs. Therefore, this work presents a procedure for the modeling of active holonic fault-tolerant control system (AHFTCS) that considers AFTCS' requirements and interpreted Petri net for the description of the systems' behavior. This method is applied to an intelligent building (IB) as a class of FPS and the results are presented.

Keywords: control system, system modeling, fault-tolerance, Petri net, holon, reconfiguration.

1 Introduction

Advances of mechatronic systems, communication networks and work organization methods, allied to the crescent competitiveness and the need for efficient services triggered great changes on productive systems (PSs) requiring more flexibility under different demands, such as production volume, type of product and nature of resources involved. The flexible productive systems (FPSs) were designed to attend the current production demands and the focus is in material technological transforming, the information processing and service execution. Therefore manufacturing systems as well as intelligent buildings (IBs) can be approached as FPS. These systems perform multiple and simultaneous processes, exploring the intense sharing of resources, which makes complex the supervision and control of the systems global behavior [1-3].

The supervision and control systems have been evolved from a centralized and hierachic architecture to a heterarchical and distributed architecture. This distributed system (DS) is composed of various sub-systems (that can be physically installed at

different geographical locations), in which the tasks are divided according to the required functionality and processing capacity of each equipment [3]. On the other hand, to assure that a FPS does not suffer interruption due to faults of its components, an AFTCS (*active fault-tolerant control system*) mechanism must be considered [4]. This mechanism involve the detection of the fault, study of its effects, identification of causes and finally, the system reconfiguration that is done by relocating processes and choosing alternative interaction paths between processes [5]. In case of faults the strategy is to recover the system functionalities (regeneration) or to maintain critical operations in such a way that some parts of the system are disabled, but not affecting other parts of the system (degeneration).

In this context, the integration of MAS (multi-agent system) and HS (holonic system) techniques with mechatronic technology, called holonic control system (HCS), is considered a trend for the intelligent automation of PSs [1, 3, 6, 7]. The aim is to explore MAS and HS concepts of superposition, such as autonomy, reactivity, proactivity, cooperation, social capacity (i.e., consideration of the human interaction on processes), and learning resources; and to take advantage from the complementary features in the implementation of HSs by means of the MASs.

However, most of supervision and control systems do not adopt HCS and AFTCS mechanisms. In fact, the amount of material published about modeling of processes that consider the use of these techniques is very little [1, 4, 6]. Therefore, this research presents a procedure for modeling and operation of active holonic fault-tolerant control system (AHFTCS) considering a sustainability approach and its functional specifications in normal circumstances and also during faults. The application of a HVAC (heating, ventilation and air conditioning) subsystem of an intelligent building (IB) [8] is presented to illustrate the advantages of the procedure.

2 Contribution to Technological Innovation for Sustainability

The new strategic approach “sustainability research” addresses the three conflicting aspects: contributing to economic development, being ecologically acceptable and socially just [9]. It is expected that this approach will lead to more sustainable solutions and thus more effective and efficient spending of public and private funds for research and development. To live “sustainability” and make this a brand for the 21st century requires a strong engagement of science, industry and politics. Priority topics of relevance suggested by the participants for the German-Brazilian cooperation on science for sustainability were highlighted during the discussion and included: renewable energy, transportation and logistics, environmental technologies, sustainability in buildings, especially governmental buildings and industrial plants, and others [9, 10].

In this context, the rational use of energy, minimization of operational costs, larger safety and comfort to the users are essential characteristics in intelligent buildings (IBs). But as mentioned before, IB can be approached as a case of FPS. That is, this work is also a contribution for the technological innovation in design methods of IBs.

Considering also previous works in the area of IBs and FPSs, here we adopt the approach of discrete event system (DES) [3, 8], i.e. Petri net (PN) and its extensions is used for description of the system behavior (its productive processes). If compared to

other description techniques of DES, PN has an equivalent modeling power and it also has the characteristic and advantage of system visualization [11, 12].

A survey [1-4, 6, 7, 15] shows that: i) there is a small number of works that consider the integration of HCS and AFTCS requirements; ii) there are few practical applications for these agent technologies, showing that there is still a long way to go to spread these HSs, iii) in most of these systems, there is no negotiation mechanism between holons, iv) there is no information about the use of a systematic method to structure and rationalize the proposed development models, since phase of specifications until operation one, such as PN models can be used.

Therefore, to model the dynamic behavior of FPS, a place/transition Petri net class was adopted, herein called extended Petri net (e-PN), to which temporized transitions, inhibitor arcs and enabling arcs (terms related to PN are in Arial) were added [11]. To systematize and make easier the modeling these models a channel/agent PN type called PFS (Production Flow Schema) [11] is used. The system's dynamic models are generated by means of e-PN. Thus, the procedure combines the bottom-up approach and the top-down approach of the stepwise refinement associated to PFS.

According to Fig.1a, the procedure presents mechanisms that allow the switching of the control in two modes: hierachic and heterarchic control architecture. It allows the switching of control between two operational modes: the “stationary mode” where the control system is coordinated in a hierarchical; and the “transient mode”, where to assure more system flexibility and agile behavior. This architecture is described in Section 3.

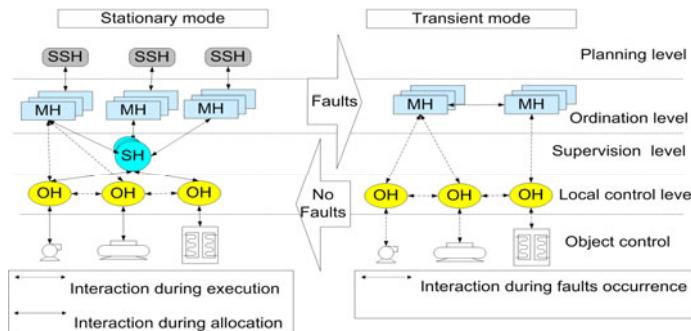


Fig. 1. Control Architecture of AHFTCS

The procedure adapts what is presented in [13] to specification of a mechanism called “diagnoser” using an e-PN model. Toward decision making phase of the AFTCS, some inference rules based on reasoning [14] may be adopted for the specification of a mechanism called “decider”. Besides, a system that considers the reconfiguration requires redundant resources to keep an adequate performance, and must also consider the transmission of control signals as part of the system to be controlled, because a fault on this communication network may also limit the coherence of command actions [4, 15].

3 Procedure for AHFTCS Design

Here the basic structure of the AHFTCS development procedure is presented: analysis of requirements, modeling, analyzing/simulation, implementation and operation. In the following explanation of each phase of the procedure it is presented some examples of models derived from case study applied. In Fig. 2a, is presented a cold production subsystem from a commercial building in São Paulo city, Brazil, composed of two chillers, four pumps, two block valves and one flowmeter (F).

Phase 1 – analysis of requirements – on this phase is defined the AHFTCS' specifications: aim of the system, control object, control devices, definition of tasks, strategies and control functions, and description of the interaction between the parts of the system, and the cases of reconfiguration.

Sub-phase 1.1 – identification of holons – on this sub-phase the holons are identified, i.e., SSH, MH, SH and OH. The holarchies are represented by ellipsis and one holon may belong, simultaneously, to various holarchies (Fig. 2b). The identification of SSH – subsystem holon – involves the definition of control functions of each product/service offered by the FPS' subsystems and how to perform production/service orders. Thus, SSH contains all knowledge necessary to operate the FPS and to choose the better strategy to reach the objectives planned. The MHs – manager holons – are the entities responsible for the management of control strategies that must be followed during execution phase. The SHs are responsible for coordinating the OHs. The SH – supervisor holon – contains all knowledge necessary to coordinate holons on lower hierachic levels. The function of the SH involves the preparation of a program of tasks and coordination of decisions for the performance of these tasks. When a process requests a resource, in fact it is requesting functionality and the SHs check the available resources to control the allocation of the resource. The OH – operational holon – represents human operators and plant's physical resources, which have any control device for its operation and establish these resources' behavior according to the objectives and skills. OH manages the behavior of these resources according to the objectives, characteristics and skills. According to Fig. 2, a holarchy (CP) is formed by the SH cold production controller (SH CP) and other holarchies: main production (MP), auxiliary production (AP) and distribution (Dist) and these holarchies are represented at other SH and OHs. For this subsystem, which provide cold water to the heat exchanger of the air conditioning unit; the control actions are developed considering redundancy: the activity [production of cold water] may be carried out in the main cold subsystem, in the auxiliary cold subsystem, or both.

Sub-phase 1.2 – AFTCSs specifications – in this sub-phase is identified the main critical points of the system and the faults that may affect the normal performance of functions indispensable to the system. After that identification it is necessary to analyze which critical processes will be subject to reconfiguration. The functions of the AFTCS are divided into four phases and are present at each holon independently of the type. The “estimation” phase involves: 1) detection of symptoms that may supervise the existence of faults and 2) the isolation of the fault. When the symptoms detected do not allow any conclusion, the system must be programmed to identify the kind of fault detected in similar cases or request external intervention. The “planning” phase decides upon the reconfiguration action based on pre-defined priorities such as:

lower performance fall, lower recovery time, etc. The “execution” phase involves the sending of commands for the performance of the selected action plan. The last phase is “learning”, which involves the storage of the relevant data in relation to the performed plan. Therefore, it may be stated that AHFTCS acts according to the following AFTCS rules: if *<symptoms>* then *<selects fault>*; if *<fault selected>* then *<selects action>*; if *<action selected>* then *<activates reconfiguration>*; and if *<reconfiguration performed>* then *<storage relevant data>*.

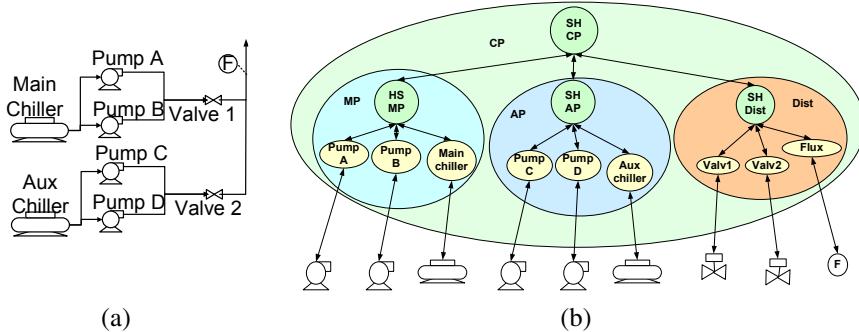


Fig. 2. (a) Cold production sub-system and (b) organization holarchies

Sub-phase 1.3 – definition of interaction patterns between holons – three interactive processes are considered in this sub-phase: "request for products/services", "execution" of products/services, "fault treatment" and "reconfiguration" due to faults. The synchronization of e-PN models is made by enabling arcs and inhibitors (Fig. 3). These interactions are extracted from UML sequence diagram [16].

Phase 2 – modeling considering reconfiguration – using PFS models represents the interactions of negotiation between holons, and the submission of orders to operational holons OHs; preparation and performance of these orders; and the treatment of faults upon their occurrence. The occurrence of faults must be represented by means of SHs and OHs' models. The control strategies of the AFTCS are modeled on this phase, with the “diagnoser” and the “decider” to fulfill the requirements of the diagnosis and decision phases. The steps to design the e-PN model of the “diagnoser” are: i) construction of e-PN models for the components of the control object; ii) construction of e-PN models of control strategies; iii) definition of observable events – generally those related to control strategy commands; and non-observable events [13], generally related to faults; iv) construction of e-PN models of sensors; v) initiate the construction of the “diagnoser” from the initial state considered “normal” (without faults); vi) relate, by means of transitions and enabling arches, the performed strategies with possible observable and non-observable events which may happen from the initial state; and vii) relate the states obtained with the states of the sensors. If the “diagnoser” does not indicate the correct state then the possible faults’ causes must be inferred to solve possible conflicts. This decision mechanism is called “decider” and its decision making rules may be based on probabilistic data, for example. Figure 3 also shows the valve 1 component commanded by the OH valve 1; the PFS and e-PN models of valve 1 model considering the influence of the control

signal transmission network; the diagnoser for the valve 1 and flowmeter; and the related decider device. In Fig. 4 is presented an example of fault treatment and its reconfiguration (degeneration).

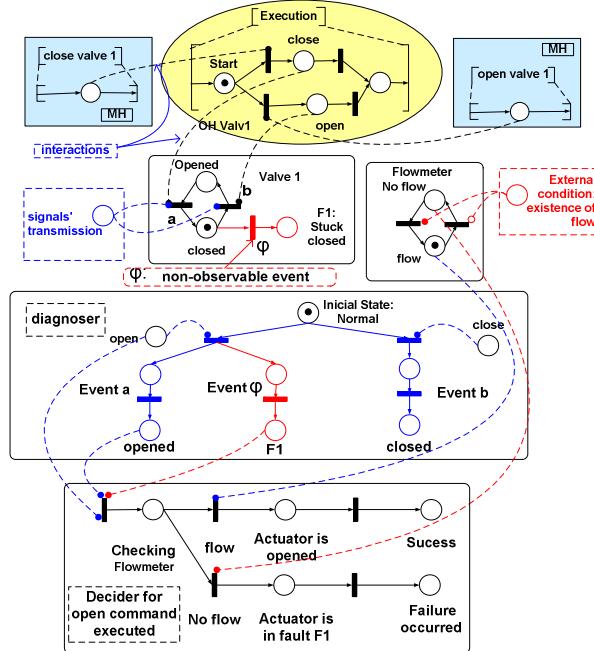


Fig. 3. Example of interactions, and control objects models, diagnoser and decider

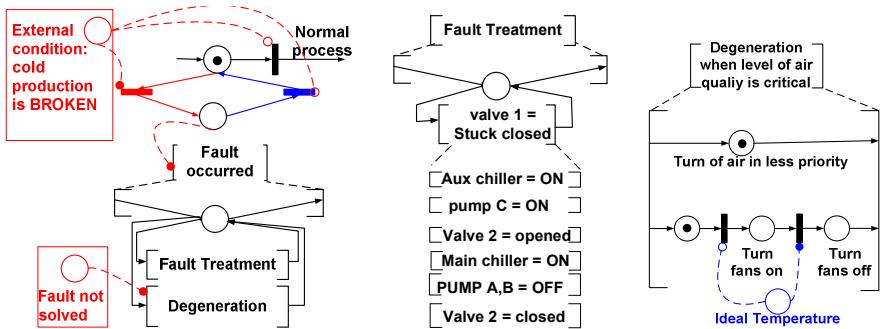


Fig. 4. Example of fault treatment and its reconfiguration (degeneration)

Phase 3 – analysis/ simulation – the analysis is developed with e-PN tools for edition and structural analysis. The behavior and the quantitative analysis were carried out by means of associated simulation techniques with checking of e-PN properties. This type of analysis allows re-design and re-engineering of the control system during the design phase. This phase is subdivided in: qualitative and

quantitative analysis. Qualitative analysis allows the verification of structural properties and behavioral models, sketching conclusions about the system operation, such as: i) liveness, that is related to the complete absence of deadlocks in operating systems; ii) reachability, to study the dynamic properties of any system; iv) reversibility, to recover from disruptive events of the operation; v) conservation and boundedness to verify the variation of the number of tokens of the net. The quantitative analysis requires the introduction of the time parameter associated with the transitions. Thus, it is possible check if the firing is consistent with specifications of the models.

Phase 4 – implementation – for the practical use, the resulting models are interpreted as control program specifications to be performed by computers (supervisory control) and programmable controllers (local control level). This phase also comprises the codification, parameterization and development of wrapper interfaces.

Phase 5 – operation – in this phase, the real-time supervision of the automation control system is performed by synchronizing the operation of the AHFTCS with the e-PN models, in order to control and monitor the system. The signals from the sensors and the status of mechatronic devices are acquired and connected with e-PN models. The adaptation and re-configuration of the FPS is supported using this procedure, i.e., the introduction or remotion of new components requires the addition or remotion of a new token in the corresponding e-PN models and, in some cases, the modification of associated holons models.

4 Conclusions and Future Works

Using as application example an intelligent building (IB), a novel procedure for design of AHFTCS considering normal operations and occurrence of faults in flexible productive system (FPS) was presented. The process combines the requirements of the holonic control system (HCS) and AFTCS (active fault-tolerant control system), with special attention to the system's reconfiguration. The modeling process is based on interpreted Petri net (PN), and its extension called PFS is used to structure the development of components' models and presentation of the proposed procedure, combining the bottom-up approach and the top-down approach of the stepwise refinement associated to PFS. The use of this systematic technique, to structure and rationalize the models development of the proposed architecture allows an environment that facilitates the development of new models. The proposed architecture and its mechanisms allow implementing a hierachic or heterarchic control structure and reacts to faults more agilely. This work synthesizes Silva's project [3], which involved modeling of the whole HVAC and other subsystems of IB such as: access control, fire fight and prevention, people transportation/ movement, and signals transmission control; besides the simulation and validation of extended Petri net (e-PN) models.

The PhD thesis of one of the authors, the student Silva, involves the whole life cycle of automation systems. Since the research done so far does not yet offers a complete solution to extending the research results towards applicability in other supervision and control systems. More detailed case studies for a complete evaluation are needed. The survey of theories, tools and applications, are considered as the most feasible and adequate research strategy in this study. The next stage of research

involves the transformation of a conceptual model, which must be developed and refined by the general surveys, to a practical model.

References

1. Schoop, R., Colombo, A.W., Suessmann, B., Neubert, R.: Industrial experiences, trends and future requirements on agent-based intelligent automation. In: Proceedings of IECON the Annual Conference of IEEE Industrial Electronics Society, Seville (2002)
2. Leitão, P., Colombo, A.W.: Petri net based Methodology for the Development of Collaborative Production Systems. In: Proceedings of the 11th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA 2006), Prague, pp. 819–826 (2006)
3. Silva, R.M., Arakaki, J., Miyagi, P.E., Junqueira, F., Santos Filho, D.J.: Intelligent Building - Modeling and Reconfiguration using Petri net and Holons. In: Proceedings of ICNPAA: the 8th IEEE Int. Conf. on Mathematical Problems in Engineering, Aerospace and Sciences, São José dos Campos, Brazil (2010)
4. Zhang, Y., Jiang, J.: Bibliographical review on reconfigurable fault-tolerant control systems. *Annual Reviews in Control* 32, 229–252 (2008)
5. Arakaki, J., Miyagi, P.E.: Degeneration methods in intelligent building control system design, Boston. IFIP, vol. 220, pp. 469–478 (2006)
6. Sousa, P., Ramos, C., Neves, J.: The Fabricare System. *Production Planning & Control* 15(2), 156–165 (2004)
7. Colombo, A.W., Neubert, R., Schoop, R.: A solution to holonic control systems. In: Proceedings of ETFA the 8th IEEE Int. Conf on Emerging Technologies and Factory Automation, Sophia/Nice (2001)
8. Wong, J.K.W., Li, H.: Construction, application and validation of selection evaluation model (SEM) for intelligent HVAC control system. *Automation in Construction* 19, 261–269 (2010)
9. Zickler, A., Mennicken, L.: Science for Sustainability: The Potential for German-Brazilian Cooperation on sustainability-oriented Research and Innovation – Introduction. In: Proceedings of the 1st German-Brazilian Conference on Research for Sustainability, São Paulo, Brazil (2009)
10. Blackstock, K.L., Kellyb, G.J., Horseyb, B.L.: Developing and applying a framework to evaluate participatory research for sustainability. *Ecological Economics* (60), 726–742 (2007)
11. David, R., Alla, H.: Petri nets for modeling of dynamic systems – a survey. *Automatica* 30(2), 175–201 (1994)
12. Hasegawa, K., Miyagi, P.E., Santos Filho, D.J., Takahashi, K., Ma, L.Q., Sugisawa, M.: On resource arcs for Petri net modeling of complex shared resource systems. *Journal of Intelligent & Robotic Systems* 26(3/4), 423–437 (1999)
13. Sampath, M., Sengupta, R., Lafortune, S., Sinnamhoideen, K., Teneketzis, D.C.: Failure diagnosis using discrete-event models. *IEEE Trans. on Control Systems Technology* 4(2), 105–124 (1996)
14. Kuipers, B.: Qualitative Reasoning: Modeling and Simulation with Incomplete Knowledge. The MIT Press, Cambridge (1994)
15. Scheidt, D.H.: Intelligent Agent-Based Control. Johns Hopkins APL Technical Digest 23(4), 383–395 (2002)
16. Booch, G., Rumbaugh, J., Jacobson, I.: The Unified Modeling Language User Guide. Addison Wesley Longman, Inc., Amsterdam (1999)

Design of Supervisory Control System for Ventricular Assist Device

André Cavalheiro¹, Diolino Santos Fo.¹, Aron Andrade², José Roberto Cardoso¹, Eduardo Bock², Jeison Fonseca², and Paulo Eigi Miyagi¹

¹ Department of Mechatronics and Mechanical Systems Engineering,
Escola Politécnica da USP, São Paulo, Brazil

² Department of Bioengineering, Institute Dante Pazzanese of Cardiology, São Paulo, Brazil

Abstract. When a patient have severe heart diseases, Ventricular Assist Device (VAD) implantation may be necessary. However, the improvement of the interaction between the device and the patient's behavior is crucial. Currently, the control of these pumps does not follow changes in patient behavior and the devices are no safe. Therefore, if VAD has no faults tolerance and no dynamic behavior according to the cardiovascular system performance, there is a serious limitation on expected results. This research investigates a mechatronic approach for this class of devices based on advanced techniques for control, instrumentation and automation to define a method for developing a hierarchical supervisory control system to control a VAD dynamically and securely. To apply this method, concepts based on Petri nets and Safety Instrumented Systems are used. This innovation reduces the interventions and unnecessary drugs, enabling a reduction of disposable material and patient hospitalization, and contributes to sustainability concept..

Keywords: Ventricular Assist Device, Petri nets, Safety Instrumented System.

1 Introduction

The increase of resources consumption on the world is one of the main features that justify the use of sustainability concept on methods for projects development. In this sense, automation can help to optimize the resources utilization and to improve devices performance. Thus, this work proposes a development of Ventricular Assist Device (VAD) to aid patient with heart failure to be able to have relatively normal life despite the disease applying sustainability concepts. This device can be used in several cases during the period that patient is waiting for heart transplantation, during a pre or postoperative recovery period, or as destination therapy when the patient has no indication for heart transplantation due to several reasons [1, 2, 3]. VAD projects involve many research areas: mechanical and electromechanical engineer, biomaterial, medicine, and also computer technologies for data collection, processing and making decision, therefore, sensors to indicate blood pressure, blood flow, body temperature and cardiac frequency are necessary [4]. Figure 1 represents how the VAD interacts with the cardiovascular system.

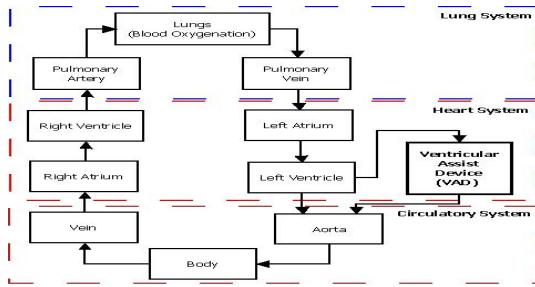


Fig. 1. VAD and cardiovascular system

In this context, two main aspects must be considered:

- First: device shall demonstrate correct and accurate performance; otherwise, if the pump fail during operation and there is no embedded system that enables the treatment of faults autonomously, serious risks to patient will be inevitable.
- Second: many VADs maintain constant blood flow regardless to patient daily needs [5], ie, they help blood circulation and do not react properly to changes [6].

Therefore, if VAD has no faults tolerance and no dynamic behavior according to the cardiovascular system performance serious limitation on results from this application may be observed. This work proposes application of a mechatronic approach based on advanced techniques for control, instrumentation and automation. These techniques allow treatment of fundamental limitations of current solutions. So, we propose a method for specifying a supervisory control system for a VAD that:

- Specify a logic for pump speed control, according to patient's dynamic behavior. Models based on Bayesian networks (BN) [12] should be applied to diagnose patient's dynamic state, at every moment and to act in VAD control.
- Specify a logic for safety interlock to prevent faults in VAD. We must diagnose the critical states by using BN and implement a diagnosis system by using Petri nets (PN). Once implemented the diagnosis control system, in parallel, should be implemented faults treatment according to specification of safety instrumented functions (SIFs) [10, 11]. These functions must be modeled in PN [9] for generating the control algorithm for faults treatment.
- Check the mathematical model of supervisory control system according to its interaction with a model of human cardiovascular system [16, 17, 18].

So, supervisory control system can be implemented and specified for *in vitro* and *in vivo* testing in a consistent way.

2 Technological Innovation and Sustainability

There are three conflicting aspects that are considered if a new strategic approach as “Sustainability research” is adopted, i.e., contributing to economic development, being ecologically acceptable and socially just [20]. This approach will guide

sustainable solutions that have to be more effective and efficient spending of public and private funds for research and development. To do the sustainability and make this a brand for the 21st century requires a strong engagement of science, industry and politics.

Considering also previous works in the area of VADs, here we adopt the approach of a hybrid system presents dynamical behavior in which simultaneous evolution of continuous and discrete state variables occurs, ie, Continuous Variable Systems (CVS) behavior merge with Discrete Event Systems (DES) behavior [8]. Considering the requirement that a VAD needs to perform control functions to adjust pump speed according to changes on cardiac frequency and needs to react against occurrence of critical faults, we proposed a hybrid supervisory control system [7].

In this context, this research will contribute to add to VAD the following features: rational use of energy, minimization of operational costs and larger safety and comfort to the users. Then, this work is also a contribution for the technological innovation in design methods of VADs.

3 Materials and Methods

Production Flow Schema (PFS) is a technique that can be used to model the set of activities that VAD can perform. The PFS is a bipartite graph composed of activity elements (action, execution), distributing elements (collect, accumulate and/or store information or items) and oriented arcs to connect the elements. Figure 2(a) shows the graphical representation of these elements.

Details of each activity modeled in PFS can be refined using PN which are capable of representing dynamic behavior of device. As VAD has continuous variables, Hybrid Petri Nets (HPN) are necessary.

HPN model has been introduced as extension of discrete PN model been able to handle real numbers in continuous way and allowing us to express explicitly the relationship between continuous values and discrete values while keeping good characteristics of discrete PN soundly. In HPN model, two kinds of places and transitions are used: discrete/continuous places and discrete/continuous transitions. A continuous place holds a nonnegative real number as content. A continuous transition fires continuously in the HPN model and its firing speed is given as function on model places. Figure 2(b) shows graphical notations of HPN elements [9]. The refinement of a model generated in PFS for a model in HPN is made based on the procedure adopted in Villani [7].



Fig. 2. (a) Production Flow Schema elements; (b) Basic elements of Hybrid Petri Net [9]

For modeling of diagnosis and treatment of critical faults, we are using the concept of Safety Instrumented System (SIS) and BN. According to Squillante [10] SIS is a layer of control in order to mitigate the risk or taking the process in a safe state. Definition of faults is made from the identification of Safety Instrumented Functions (SIF). In this way, a SIF describes a critical fault that should be diagnosed and treated by SIS. A SIS implements its SIFs through sensors and devices perform control by actuators. For each SIF a parameter called Safety Integrity Level (SIL) is defined. This parameter is a measurement of safety for components and/or systems. SIL reflects what end users can expect from a device and/or system in a safety function, and in case of fault, this occurs in a safety way. BNs provide formalism for reasoning about partial beliefs under uncertainty conditions. The propositions are as numerical parameters signifying the belief degree according to some evidence or knowledge. So, formally, BNs $B = (G; Pr)$ are made up by a topological structure G and a set of parameters Pr that represents probabilistic relationship among their variables [10, 11].

Therefore, to develop a design for supervisory control system for VAD, according to the modeling techniques presented above, proposes a method for developing the VAD a supervisory system shown in Figure 3:

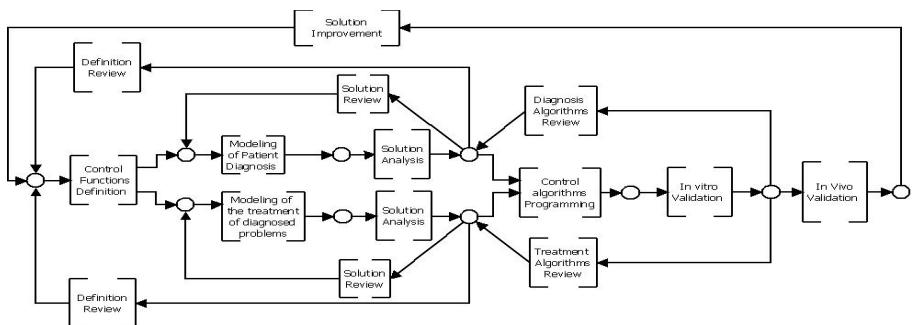


Fig. 3. Method for supervisory control system design applied to VAD

The sequence proposed aims to organize the activities involved to develop control system design applied to VAD. So, these activities (represented by the elements of the activity model in the scheme proposed PFS) are presented briefly:

- Definition of control functions - A team of doctors and engineers defines the degree of autonomy of VAD control system. The team responsible control system development needs to select ideas that are possible to be implemented taking into account: sensors available, VAD performance characteristics and technological limitations. At this stage, a table with VAD control functions is specified.
- Modeling of Patient Diagnosis – Initially, diagnosis model involves development of a cause and effect matrix, which is basis for a BN. Then, this network can be converted into a HPN model.
- Modeling of treatment for diagnosed problems - From HAZOP (hazard and operability) [13] study the risk analysis report from VAD is obtained. Based on this information, we have the SIF, SIL and events (from sensors) and actions (to

actuators) for each SIF. Next, diagnostic model and corresponding SIFs are modeled in HPN for control system design.

- Solution Analysis – First, structural analysis of HPN model of diagnostic process is made. Next, it is checked if the HPN has no deadlock states (markings where no transition is enabled). For at, a simulator [14] can be used.
- Control algorithm programming - A Programmable Controller (PC) is an essential equipment for implementation of control systems. Consequently, standards have been set for this equipment allowing the reuse of existing software modules and ensuring high quality solutions, especially for conditions that require security methods for verification and validation. Thus, it is necessary to adopt following procedure: generation control program in programming language according to IEC 61131-3 [15], based on conversion of HPN models.
- In *vitro* validation for control algorithm - Control algorithms can be validated using mathematical model that simulates human cardiovascular system. The entire electrical equivalent of the model is shown in Figure 4. The electronic parameters are correlated to their mechanical parameters as follows: voltage (volt) is analogous to pressure (mmHg), capacitance (F) to compliance (ml/Pa), resistance (Ω) to resistance (1 Pa.s/ml), and inductance (H) to inertance (1 Pa.s²/ml) [16]. The elements of each artery including one or two resistor, an inducer and a capacitor. Figure 4(a) belongs to the arteries with the radius of less than 0.2 cm and figure 4(b) belongs to the rest of the arteries. The architecture used to validate VAD control with cardiovascular system presented in Figure 1. Next step is prototype implementation that can be validated by a simulator of cardiovascular system. Currently, the Institute Dante Pazzanese of Cardiology (IDPC) has a programmable mechanical simulator that performs in *vitro* testing and is able to simulate real situations that can occur in patient's behavior [19].

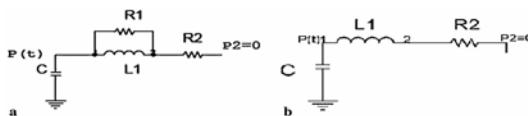


Fig. 4. (a) Electronic elements equivalents arteries with radius less 0.2 cm (b) others arteries

- In *vivo* validation for control algorithm - After simulation and in *vitro* validation of control algorithm, VAD is ready for in *vivo* tests in calves [2].

Applying this set of procedures, is definition of a method for supervisory control system design is possible and capable to provide changes in VAD rotation speed, according to changes in cardiac frequency of patient, and can improve security and quality of life for patient who needs this type of device.

4 Discussion of Results and Contributions

The project of the control system of VAD according to the procedure previously considers: (i) critical faults from HAZOP study for VAD, (ii) BN for diagnosis and decision, (iii) definition of Safety Instrumented Functions (SIF) using HPN and (iv)

modeling of supervisory control system considering discrete and continuous variables of VAD. The result is the logical ordering of supervisory control system functions that are shown in Figure 3 based on PFS formalism (Figure 5). To implement this control system for VAD at IDPC is proposed the architecture according to Figure 6.

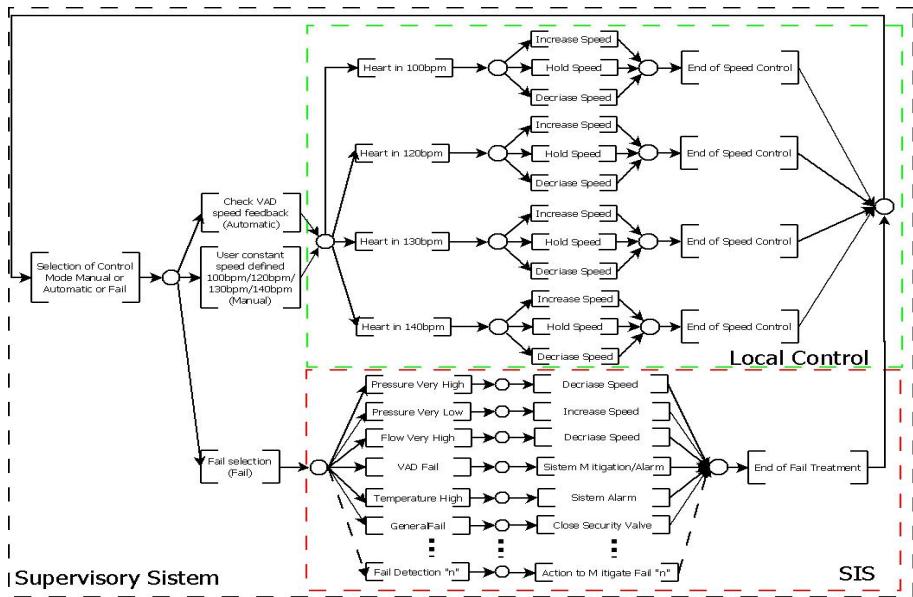


Fig. 5. PFS model to VAD supervisory control system

According to this work proposal, each activity of model presented at PFS of Figure 5 is represented by a place on HPNs, and the marks can represent local and global states of VAD control system. Transitions are synchronized with human body reactions through adequate sensors. Oriented arcs can define the sequence for control functions processing. Therefore, we apply obtain VAD control algorithm based on proposed method in the proposal control architecture as show in Figure 6.

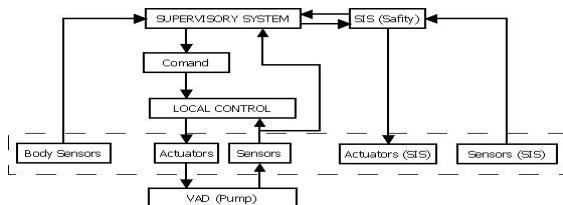


Fig. 6. Control architecture proposed

Therefore, the control system proposed can improve VAD and can fit patient needs providing better patient quality of life and longer survival. Supervisory system can also assist in diagnosis and in interventions to maintain VAD functions.

5 Conclusion

The VAD under development at IDPC presents difficulties concerning blood flow adjustment according to patient state: there is no device supervisory control that can adjust rotation speed based on patient needs and there is no treatment of VAD faults that can help patient's safety.

With a control system automatic and dynamic is possible that VAD adjust patient needs, providing a higher quality of life and enabling a patient survival. The supervisory system can also assist in diagnosis and possible interventions that doctors need to VAD control. Thus, VAD system may be adapted to the patient needs, keeping security and provides risks reduction. Therefore, VAD supervisory control system proposed offers advantages compared to currents VADs control systems.

This study contributes to have a customized VAD considering the patient's illness and different metabolism. Moreover, Safety Instrumented System concept is essential to provide risk reduction that might interfere in VAD functioning and in patient's life.

5.1 Future Works

About the method to obtain VAD control architecture:

- Make research about a most efficient method for setting requirements to improve control system autonomy, concerning items complexity to determine the control and security system and to optimize process to obtain supervisory system.

About sensors and actuators used:

- Work to improve sensors to make them less invasive and allow to provide signals with higher quality and precision;
- Work to improve the dynamic pump features to improve system efficiency.
- Add block valves to VAD applying new technologies and using biomaterials.

Acknowledgments. We are grateful support of FAPESP and MEC/CAPES/PET.

References

1. Wada, E.A.E., Andrade, A.J.P., Nicolosi, D.E.C., Bock, E.G.P., Fonseca, J.W.G., Leme, J., Dinkhuyzen, J.J., Biscegli, J.F.: Review of the spiral pump performance test, during cardiopulmonary bypass, in 43 patients. ASAIO Journal 1, 1–2 (2005)
2. Andrade, A.J.P., Nicolosi, D.E.C., Lucchi, J.C., Biscegli, J.F., Arruda, A.C.F., Ohashi, Y., Muller, J., Tayama, E., Glueck, J., Nosé, Y.: Auxiliary total artificial heart: A compact electromechanical artificial heart working simultaneously with the natural heart. Artificial Organs 23, 876–880 (1999)
3. Fonseca, J.W.G., Andrade, A.J.P., Bock, E.G.P., Leme, J., Dinkhuyzen, J.J., Paulista, P.P., Manrique, R., Paulista, P.H., Valente, P., Nicolosi, D.E.C., Biscegli, J.F.: In vivo tests with the auxiliary Total Artificial Heart as a left ventricular assist device in calves. ASAIO Journal 1, 1–5 (2005)
4. Ohashi, Y., Andrade, A.J.P., Muller, J., Nosé, Y.: Control System Modification of an Electromechanical Pulsatile Total Artificial Heart. Artificial Organs 21(12) (1997)

5. Fonseca, J.W.G., Andrade, A.J.P., Nicolosi, D.E.C., Biscegli, J.F., Legendre, D.F., Bock, E.G.P., Lucchi, J.C.: A New Technique to Control Brushless Motor for Blood Pump Application. *Artificial Organs* 32 (2008)
6. Bock, E.G.P., Ribeiro, A.A., Silva, M., Antunes, P.I.T.C., Fonseca, J.W.G., Legendre, D.F., Leme, J., Arruda, A.C.F., Biscegli, J.F., Nicolosi, D.E.C., Andrade, A.J.P.: New Centrifugal Blood Pump With Dual Impeller and Double Pivot Bearing System: Wear Evaluation in Bearing System, Performance Tests, and Preliminary Hemolysis Tests. *Artificial Organs* 32 (2008)
7. Villani, E., Miyagi, P.E., Valette, R.: Landing system validation based on Petri nets and a hybrid approach. *IEEE Transactions on Aerospace and Electronic Systems* 42(3/4), 1420–1436 (2006)
8. Ho, Y.: Discrete Event Dynamic Systems: Analysing Complexity and Performance in the Modern World. IEEE Press, Los Alamitos (1992)
9. Matsuno, H., Tanaka, Y., Aoshima, H., Doi, A., Matsui, M., Miyano, S.: Biopathways representation and simulation on hybrid functional Petri net. In: *Silico Biol.* (2003)
10. Squillante, R., et al.: Safety instrumented system designed based on Bayesian network and Petri net. In: 8th Intern. Conf. on Mathematical problems in Engineering, Aerospace and Sciences (ICNPAA), São José dos Campos, Brazil (2010)
11. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufman Publishers, San Mateo (1988)
12. Cooper, G.F., Herskovits, E.: A Bayesian method for the induction of probabilistic networks from data. *Machine Learning* 9, 309–347 (1992)
13. IEC, International Electrotechnical Commission, Functional Safety of Electrical / Electronic / Programmable Electronic Safety-related Systems (IEC 61508), Geneva, Switzerland (1998)
14. Visial Object Net; Petri Net based Engineer Tool; version 2.7a; Copyright Dr. Rainer Drath (2007), <http://www.paramsoft.de/>
15. IEC, International Electrotechnical Commission, Programmable Controllers Part 3, Programming Languages, IEC1131-3. Geneva: IEC, p. 207 (1993)
16. Mona, A., Mahdi, N., Kamran, H.: Mathematical Modelling and Electrical Analog Equivalent of the Human Cardiovascular System. *Cardiovasc. Eng.* (2010)
17. Reed, T.R., Reed, N.E., Fritzson, P.: Heart sound analysis for symptom detection and computer-aided diagnosis. *Simulation Modelling Practice and Theory*, 129–146 (2004)
18. Sainte, M.J., Chapelle, D., Cimrman, R., Sorine, M.: Modeling and estimation of the cardiac electromechanical activity. *Computers and Structures*, 1743–1759 (2006)
19. Andrade, A.J.P., Filipini, C.L., Lucchi, J.C., Fonseca, J.W.G., Nicolosi, D.E.C.: An electro-fluid-dynamic simulator for the cardiovascular system. *Artificial Organs* 32 (2008)
20. Zickler, A., Mennicken, L.: Science for Sustainability: The Potential for German-Brazilian Cooperation on sustainability-oriented Research and Innovation – Introduction. In: Proceedings of the 1st German-Brazilian Conference on Research for Sustainability, São Paulo, Brazil (2009)

Multi-agent Topologies over WSANs in the Context of Fault Tolerant Supervision

Gonçalo Nunes^{1,3}, Alberto Cardoso¹, Amâncio Santos^{1,2}, and Paulo Gil^{1,3}

¹ CISUC, Department of Informatics Engineering, University of Coimbra, Portugal

² Instituto Superior de Engenharia de Coimbra, Portugal

³ Departamento de Engenharia Electrotécnica, Faculdade de Ciências e Tecnologia
Universidade Nova de Lisboa, Portugal

{gnunes, alberto}@dei.uc.pt, amancio@isec.pt, psg@fct.unl.pt

Abstract. Wireless Sensor and Actuator Networks can be used to detect and classify ephemeral distributed events, where different process components with different behaviours are involved. Agents implementing Distributed Artificial Intelligence techniques are a key value in improving the overall system's performance. This paper proposes a general WSAN Multi-Agent based architecture for robust supervision and fault tolerant control.

Keywords: Agents, Mobile Agents, Multi-Agent Systems, Fault Detection and Supervision, Wireless Sensor and Actuator Networks.

1 Introduction

Wireless Sensor and Actuator Networks (WSANs) has attracted considerable attention in the last few years. They are distributed networks of sensors and actuators nodes, which act together in order to monitor and/or control a diversity of physical environments [1]. Each node is a small electronic device with wireless communication capabilities, including data storing and processing power, which can be programmed to interact with the physical environment by means of incorporated sensors and actuators. Additionally, they present reduced dimensions and can be used in a number of applications, including military, medical, process industry, environmental tracking, home automation, surveillance systems, just to name a few [2], [3]. In the industrial context, WSANs may be used in rare event detection or periodic data collection. In uncommon event detection, nodes are used to detect and classify rare, random, and ephemeral events, such as alarms or faults detection notifications. On the other hand, periodic data acquisition can be required for operations such as monitoring and control, reducing installation and operation costs [4], [5]. Furthermore, unlike traditional wired networks, nodes of a WSAN can be deployed in hostile or inaccessible environments, which is impractical with normal wired approaches. Because of their intrinsic features, WSAN can be a useful and powerful solution for a number of practical applications. However, since a WSAN is a preprogrammed system, it does not allow coping with unpredictable contingencies. What happens when the sensor network is faced with specific constraints, like energy consumption, data optimization, quality of service, for which it was not designed? How to achieve flexibility in the

WSAN? The answer relies to some extent on the incorporation in a single framework of distributed artificial intelligence (DAI) methodologies along with Multi-Agent Systems (MAS) [6].

Agents are intelligent and autonomous software programs capable of interacting with other software components within a given application, and sharing a common goal. The integration of agents in a given environment can be remarkable advantageous in the case of several distinct process components (exosystems) with different behaviors and dynamics, for which it is, required to communicate with each other to perform a given global task. Agents can be organized in a particular multi-agent framework, where they cooperate in order to solve particular problems, inexorably constrained to the system's teleonomy, and taking advantage of their specific skills and individual knowledge. In industrial environments they can be used, for instance, in fault diagnosis or in the implementation of reconfiguration heuristics applied to local digital controllers, or creating drivers that exhibit a certain degree of tolerance to faults.

The present work proposes a robust data processing multi-agent based framework, ensuring the performance (reliability, timeliness, and precision) and data quality monitoring (QoS - Quality of Service), as well as fault diagnosis capabilities to deal with common WSANs constraints. Section 2 introduces the concept of agent and Multi-agent based systems, describing common multi-agent topologies and their features from the perspectives of physical and software abstractions. In section 3, the multi-agent based WSAN architecture is described focusing on structural and functional issues. Finally in section 4 some conclusions are drawn.

2 Technological Innovations for Sustainability in Multi-agent Based Systems over WSANs

Agents can be defined as computing entities, comprising a certain degree of autonomy and having the ability to feel and/or actuate in order to achieve predefined goals [1], [7], [3]. Other features include *i) Autonomy*: Agents are independent entities, able to accomplish a given task, without any programming or direct intervention; *ii) Reactivity*: agents are capable of perceiving their environment and respond quickly and effectively to changes; *iii) Pro-Activity*: agents are able to take initiative goals and behave in order to meet them; *iv) Cooperation*: agents have the ability to interact and communicate with each other to meet their goals and *v) Intelligence*: in order to evaluate and take over a task in autonomous way, the agent should incorporate intelligent techniques [7], [5].

The manufacturing industry has entered an era in which computer technology has refocused attention from hardware platforms to operating systems and software components [7]. The need for continuous real-time information (available at any time to many people) is currently pushing information technology providers to develop control system models and management systems to support this need. Multi-Agent Systems offer a new approach to designing and building complex distributed systems that significantly extends previous approaches and methodologies, like object-oriented or distributed computing [8]. It promises to be a valuable software engineering solution concerning the development of complex industrial systems, where complexity and spatial distribution of processes call for new methodologies.

Committed with the industrial panorama, the WSAN appears to be a powerful solution. Although, sensor/actuator nodes have limited resources, namely, reduced physical size, small memory, limited computation, small energy budget, and narrow bandwidth [3], [13]. A large group of this small and low-cost sensor nodes can be deployed to a given domain of interest and form a distributed networking system, in which collaboration among several sensor/actuator nodes is crucial to overcome the limited sensing and processing capabilities of each node and to improve the reliability of the decision making process [6]. The concept of mobile agent is a valuable solution to deal with these challenges.

In a mobile-agent based approach [9], [11], the transfer unit is the software agent itself and is based on three intrinsic premises: *autonomy*, *communication* and *mobility*. This approach may be advantageous for collaborative information processing in sensor networks, extending the inherent functionalities of networks, allowing saving the network bandwidth and computation time, since unnecessary nodes visits and agent migrations are avoided [9]. Since their scale is very small and bandwidth requirements are reduced, the necessary energy for transmission process is very low, thereby optimizing the energy autonomy of each node [12].

3 Three-Level Architecture

This section focus on a multi-agent based data processing architecture for WSAN, guaranteeing a certain level of performance, data quality monitoring and fault diagnosis functionalities, as well as the possibility of integrating Fault Detection and Identification (FDI) techniques with Fault Tolerant Control (FTC) modules. The overall goal is to provide a suitable framework for robust supervision purposes over WSANs, to meet application specific performance targets to integrate with industry resource systems. The framework facilitates the integration of the physical environment with software agents by means of the two main agent's characteristics, namely, communication and mobility, and enables a comprehensive handling of the fault management problem by the association of each monitoring activity to a particular agent service. The multi-agent approach conceptualized in this architecture will bring to the WSAN system, the following advantages: *i) Modularity and Scalability*, instead of appending new features to a system, agents can be added/launch or deleted without breaking or interrupting the overall process or even the task currently running in a given node; *ii) Mobility*, when, in a local node, an agent thumps, it can readily be regenerated by uploading it again from the server to the local node; *iii) Reliability*, in case of an active link break-down from the local node to the supervision system, local agents could take over the closed loop system stabilization role, considering the last reference vector, until communications are re-established; *iv) Concurrency*, agents are able to perform tasks in parallel, giving more flexibility to the networked system itself and also speeding up the computation and *v) Collaborative dynamics*, agents share their resources to perform their goals.

3.1 Architecture Overview

The conceptual multi-agent based WSAN hierarchical (3-level) architecture (Figure 1) comprises the following constituting parts:

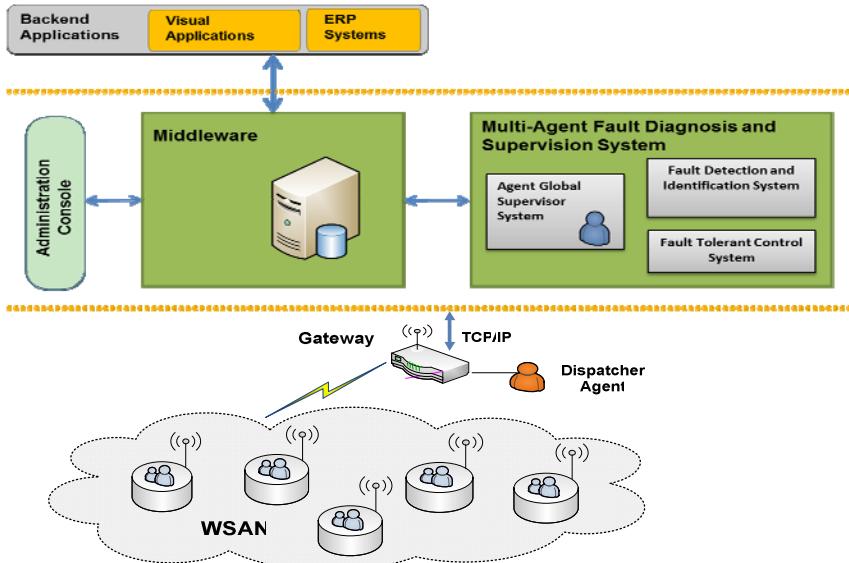


Fig. 1. Multi-agent based approach for distributed fault diagnosis and control

- i) *Middleware platform:* Middleware platforms are used to connect applications, services and components, which interact within a given system architecture. It is positioned in the middle of a typical three tier architecture. In this work an event-based middleware provides the application user with an extensive set of functionalities for developing distributed systems. It allows users to collect process and reason real-time data, as it is processed through the system. This middleware, as an event-based system, provides a strong concept of decoupling which applies to both internal middleware components (query processing and distribution or adapter framework) and external components (the WSAN, at the process level, and, at higher level, the business applications, like visual applications and ERP systems), which communicate using the middleware;
- ii) *Multi-agent fault diagnosis and supervision system:* This module is responsible for accommodating three components, namely, the Global Agent Supervisor system (GAS), which houses the agent platform responsible for the global monitoring and management of local manager agents (see section 3.3); Fault Detection and Identification system (FDI), devoted to the implementation of FDI techniques; Fault Tolerant Control module (FTC), assigned to the implementation of fault tolerant control techniques, which can be used together with the FDI system;
- iii) *Wireless Sensor and Actuator Network:* The WSAN comprises several sensor/actuator nodes, deployed in the environment for monitoring and control. Each node acquires data from exogenous systems under monitoring, to which it is assigned, this information being subsequently perceived and processed by dedicated software agents' services (see section 3.2). These local agents provide nodes with the autonomy to respond accordingly to faulty events, for instance, in case of

time-outs or latencies exceeding a predefined threshold in the wireless communications. The exogenous system outputs are converted from analogue to digital, and vice-versa, by ADCs (Analogue-to-Digital Converters) and DACs (Digital-to-Analogue Converters) that are embedded in the sensor/actuator node microcontroller (Figure 2).

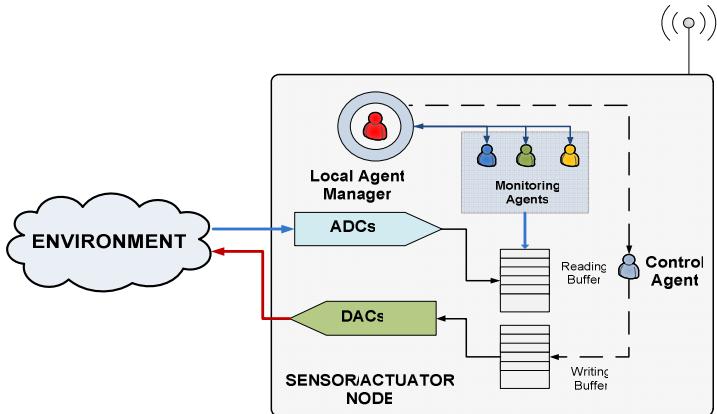


Fig. 2. Local multi-agent architecture

3.2 Local Multi-agent System

The architecture is based on the multi-agent paradigm [10], where each agent is responsible for a specific task. In terms of *modus operandi*, this framework relies on a collaborative and sharing profile/approach, which is a necessary condition to any distributed system. As can be seen in figure 2, each local sensor/actuator node includes a set of agents for monitoring and control purposes, accordingly to environment sensing and actuator constrains, as well as to trigger predefined faulty alarms to a local macro-agent (Local Agent Manager) responsible to reacts to these faulty events. The following features and services are included in each node:

i) Local Agent Manager

The main purpose of the Local Agent Manager (LAM) is to carry out extensive management routines related to other dependent (lower lever) local agents, and monitoring the communication status between the sensor node and the gateway. Whenever active, it automatically launches specialized monitoring agents, with the purposes of real-time monitoring of exogenous output.

The LAM is also responsible for continuously monitoring the status of these local subordinate agents. If one of these agents crashes, the manager kills the corresponding thread and re-launches its clone from the local agents' repository. Furthermore, in case of repeated crashes or agent's corruption, a regenerating request is sent back to the GAS, via the gateway Dispatcher Agent (DA). A default function associated with the LAM is to choose which agents should be launched and when, depending on the

environment status. If a monitoring agent is no longer necessary, the local manager removes it from memory, although it can always be re-launched whenever necessary. Another task concerns the message latency estimation, regarded as the elapsed time between the analogue data acquisition and its reception at the gateway's side using an active link. Since control actions are computed in the Fault Diagnosis and Supervision System (FDSS), in the server, and sent to the sensor node, if the communication link breaks-down the local manager should detect this faulty event and proceed accordingly, so as to not compromise the system's stability or performance. The LAM then launches a stabilizing Control Agent (CA) that is responsible to stabilize the system around the last reference received from the gateway. When the connection is re-established the local manager will remove this agent and the control authority is transferred to the FDSS.

ii) Monitoring Agents

The Local Monitoring Agents (LMAs) are launched by the local manager. By accessing the reading buffer, the LMAs continuously follows up the data read from the exogenous system or environment and is programmed to react to faulty events, such as abnormal sensor measurements or structural faults, just to name a few. In those cases, local agents are responsible for triggering alarms that are sent to the LAM. Since the computing power is limited the proposed fault diagnosis methodologies are based exclusively on univariate or multivariate statistics. In the univariate approach, the upper and lowers bounds define nominal operation conditions on the basis of a predefined threshold, and their violation triggers a predefined fault alarm. In multivariate T2 approaches more than one observation variable is used for decision-making. Let us assume we have a data set $X \in R^{n \times m}$, comprising m variables and n observations. Then, the corresponding covariance matrix is given by $S = 1/(n-1)X^T X$. Assuming that the covariance matrix is invertible, the Hotelling's T_2 statistics [14] can be given by $T_2 = z^T z$, where $z = \Lambda^{0.5} V^T x$, being Λ and V given by the singular value decomposition of S . Appropriate thresholds for T_2 statistics based on the level of significance α can be determined by assuming that observations are randomly sampled from a multivariate normal distribution. Therefore, faults can be detected for observations outside the elliptical confidence region, that is $T_2 \geq T_2\alpha$, triggering in this case a predefined fault alarm.

iii) Control Agents

The main task of Control Agents (CAs) is to stabilize the exogenous system around the last reference received from the gateway, in case of communication link breakdown. For this reason, these agents are stacked in the agent's local repository and are only launched by the LAM if this scenario (fault) occurs. Their role is crucial for the autonomy of the WSAN system, allowing the system to respond dynamically to communication time-outs, and thus ensuring the operability of the system under control, in case of connection failure.

3.3 Dispatcher Agent

The network gateway is a central component to this architecture, since it establishes the interface between the low-layered information and the upper-layered modules, resolving and forwarding packets from the WSAN to the upper modules. Regarding the implementation of supervision policies and fault diagnosis purposes, it is necessary to incorporate some intelligence on this module. The Dispatcher Agent (DA) is then responsible for analyze and identify the sender (agent) of the alarm, before sending the output alarm message to the GAS.. This data processing is very crucial to the overall supervision robustness, providing the necessary information to the Multi-Agent Fault Diagnosis and Supervising System that is implemented in the upper-layer of this architecture. The DA will also receive the action events from the GAS and forwards these events to the corresponding LAMs.

3.4 Global Agent Supervisor

The Global Agent Supervisor (GAS) is in charge for supervising and managing all the WSAN agents. It is connected to a global agent database, or a catalogue, consisting of a backup of all existing agents in the WSAN nodes, which can be used to regenerate corrupted local agents, including LAMs. When a LAM is unable to regenerate a given corrupted local monitoring agent a regenerating request is sent to the GAS. This request is processed accordingly and a clone generated from the global multi-agent repository is sent to the gateway DA, which forwards it to the respective node. As a global manager, it has the authority to reconfigure all the local agents, depending on specific premisses. This is a major advantage of the distributed system, since it makes possible, over time, the reconfiguration and adaptability of nodes' functionalities and also enables the scalability of local agents' databases. Other functionality associated with the GAS is to manage different local managers for each node, by uploading, deleting or killing a given LAM, according to specific requirements.

4 Conclusions

A conceptual multi-layered and middleware-driven multi-agent architecture for WSAN applications is presented. The multi-agent approach is a valuable engineering solution to distributed systems, such as WSAN, offering the mobility, autonomy, and intelligence to sensor nodes. This is carried out through independent software modules that use sensors and actuators to interact with the environment. The proposed architecture takes into account the specificities and constraints of sensor networks, in order to reflect the specific needs of WSAN as a distributed system, in the context of faults' monitoring and fault tolerant control. The prototype of the architecture described in this paper is in development, in a real-environment test-bed, having already produced very attractive results, that will be released in future publications.

Acknowledgments. This work has been supported by the European Commission under the contract FP7-ICT-224282 (GINSENG). The author would like to acknowledge this support.

References

1. Mendes, M.J.G.C., Santos, B.M.S., da Costa, S.J.: Multi-agent Platform and Toolbox for Fault Tolerant Networked Control Systems. *Journal of Computers* 4(4), 303–310 (2009)
2. Ospina, P.A., Canola, M.A., Carranza, O.D.: Integration Model of Mobile Intelligent Agents within Wireless Sensor Networks. In: IEEE Latin-American Conference on Communications, Medellín Colombia, pp. 1–6 (2009)
3. Tirkawi, F., Fischer, S.: Generality Challenges and Approaches in WSNs. *I. J. Communications, Networks and System Sciences* 1, 1–89 (2009)
4. Low, K.S., Win, W.N.N., Meng, J.E.: Wireless Sensor Networks for Industrial Environments. In: International Conference on Computational Modeling, Control and Automation, 2005 and International Conference on Intelligent Agents, Web Technologies, and Internet Commerce, vol. 2, pp. 271–276 (2005)
5. Cerrada, M., Cardillo, J., Aguilar, J., Faneite, R.: Agents-based design for fault management systems in industrial processes. In: *Computers in Industry*, vol. 58, pp. 313–328. Elsevier, Amsterdam (2007)
6. Biswas, K.P., Qi, H., Xu, Y.: A Mobile-Agent Based Collaborative Framework for Sensor Network Applications. In: Proceedings of the Third IEEE International Conference on Mobile Ad-hoc and Sensor Systems, Vancouver, pp. 650–655 (2006)
7. Paolucci, M., Sacile, R.: Agent-based manufacturing and control system: new agile manufacturing. CRC Press LLC, Boca Raton (2005)
8. Bussman, S., Nicholas, R.J., Wooldridge, M.: Multi-Agent System for Manufacturing Control: A Design Methodology. Springer Science, Heidelberg (2004)
9. Buse, D.P., Wu, Q.H.: IP Networked-based Multi-Agent System for Industrial Automation – Information Management, Condition Monitoring and Control of Power Systems. Springer Science, Heidelberg (2007)
10. Cheyer, A., Martin, D.: The Open Agent Architecture. *Journal of Autonomous Agents and Multi-Agent Systems* 4(1), 143–148 (2001)
11. Braun, P., Wilhelm, R.: Mobile Agents-Basic Concepts, Mobility Models and The Tracy Toolkit. Elsevier Inc., Amsterdam (2005)
12. Xiang, H., Bin, L.: A Mobile-agent-based Role Switching Management Mechanism in WSN. In: The 2009 International Conference on Computational Intelligence and Software Engineering, Wuhan, China, pp. 1–4 (2009)
13. Chen, M., Kwon, T., Yuan, Y., Leung, M.C.V.: Mobile Agent Based Wireless Sensor Networks. *Journal of Computers* 1(1), 14–21 (2006)
14. Chiang, L., Russell, E., Braatz, R.: Fault Detection and Diagnosis in Industrial Systems. Springer, Heidelberg (2001)

Switched Unfalsified Multicontroller

Fernando Costa, Fernando Coito, and Luís Palma

Departamento de Engenharia Electrotécnica,
Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa,
2829-516 Caparica, Portugal
`{lbp, fvc}@fct.unl.pt`

Abstract. In this paper, we present a controller design strategy for the implementation of a multicontroller structure for single-input single-output (SISO) plants. The overall control system can be viewed as a feedback interconnection of a SISO plant, a set of candidate controllers and a switched selection scheme that supervises the switching process among the candidate controllers. The switching scheme is designed without explicit assumptions on the plant model, based on the unfalsified control concept introduced by Safonov et al. [1, 2]. A switched multicontroller structure is implemented and experimental results are presented.

Keywords: multiple model control, adaptive control, switched control.

1 Introduction

Dealing with nonlinear systems is an inherently difficult problem. As a consequence models and analysis of nonlinear systems will be less precise than for the simpler linear case. Thus, one should look for model representations and tools that utilize less precise system knowledge than the traditional approaches. This is indeed the trend in the area of intelligent control where a range of approaches, such as Fuzzy Logic, Neural Networks and Probabilistic Reasoning are being explored [3]. The current paper uses operating regime decomposition for the partitioning of the operating range of the system in order to solve modeling and control problems.

1.1 Unfalsified Switching Control

The operating regime approach leads to multiple-model or multiple controller (multiple model control – MMC) synthesis, where different local models/controllers are applied under different operating conditions, see Fig. 1. One version of the above strategy is to represent the global system as a family of smaller local regions, where the supervisory controller alters the controller according to the current local region in which the process is operating. It must be stressed that this strategy holds only if the nonlinear system can be represented as a Linear Parameter varying (LPV) system.

The switching is orchestrated by a specially designed logic that uses the measurements to assess the performance of the candidate controller currently in use and also the potential performance of alternative controllers. In performance-based supervision

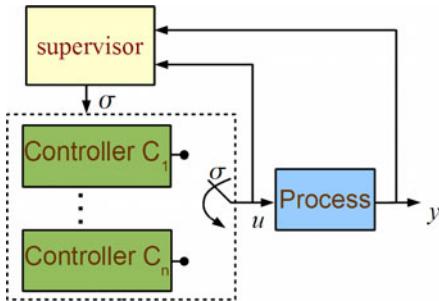


Fig. 1. Switching control. The switching decision between the controllers is performed through the switching signal σ .

the supervisor attempts to assess directly the potential performance of every candidate controller, without estimating the model of the process [1, 2, 4]. To achieve this, the supervisor computes performance signals that provide a measure of how well the controller C_i would perform in a conceptual experiment, in which the actual control signal u would be generated by C_i as a response to the measured process output y . This approach is inspired by the idea of controller unfalsification [1].

Using the unfalsification concept, no assumptions on the plant structure are required. The best controller among a set of candidate is selected straight from input/output data. The performance of all candidate controllers is evaluated directly, at every time instant, without actually inserting them in the feedback loop. Controllers that prove to be unable to drive the system according to the desired closed loop dynamics are entitled falsified. Only unfalsified controllers are candidate to actually control the process. Thus, switching between candidate controllers is based directly on their performance.

1.2 Controller Design

A key feature over the unfalsified control approach is the separation between the supervisor switching policy, and the controllers design and tuning procedure. Apart for some causality constraints, there are no relevant restrictions on individual controller structures. In fact, different controller structures may be combined into a single unfalsified switched multicontroller.

A relevant aspect on unfalsified control is the fact that, in spite no process model is required for the development of the supervisor switching scheme, all the controllers share the same closed loop specifications, usually in the form of the behavior from a reference model to be tracked. This makes its use within multi-loop control structures quite interesting, as it provides a level of decoupling between loop dynamics and the process operation conditions.

Within this framework two different approaches towards the development of such a multicontroller were developed. The first applies a set of standard state space based pole placement controllers, using Kalman filter for state estimation. The second uses non-parametric process models and the set of controllers is determined from experimental frequency response data, through frequency domain optimization. Due to space constraints this second controller will be further described in a latter article.

2 Contribution to Sustainability

The keyword that is always tied to control, even when it is not explicitly mentioned, is “performance”. Within the control field performance may be evaluated by a broad spectrum of index functions, however, in practical applications the ultimate performance assessment is related to the quality of the process outcome and the efficient use of resources – materials, energy and time. Both resources and quality are essential topics for sustainability.

3 The Unfalsified Pole Placement Multicontroller

There is vast literature on supervisory control, mainly for process estimation based schemes. Among those based in process estimation using Certainty Equivalence, interesting references are [6, 7, 8, 9], while for or Model Validation based schemes some relevant papers are [10,11]. Also, a very interesting tutorial may be found in [5] where an attempt is made to integrate the different approaches within a unified framework.

As for performance evaluation based algorithms, and specially unfalsified control, some important references are [1,2,4,12].

It is well known that switching among stabilizing controllers can easily result in an unstable system [9, 47]. To avoid a possible loss of stability caused by switching one should then require the switching logic to prevent “too much” switching, by implementing a dwell-time strategy [13, 4].

3.1 Unfalsified Controllers

Consider that the process to be controlled is unknown and that the only available information is the past values from the set-point (r), the output (y) and the control action (u). The aim is to determine if a controller is capable to lead the closed loop system to behave according to some predefined reference model W_m .

It is assumed that there is a number of predesigned “causally-left-invertible” controllers C_i (in the sense that the current value of $r_i(t)$ is uniquely determined by past values of $u(t)$ and $y(t)$), among which at least one is able to fulfill the specification. After Safonov [1] performance criterion (1) is used to evaluate discrete time controllers.

$$V(\tilde{r}, u, \tilde{e}, t) = \begin{cases} \frac{\|\tilde{e}\|_t + \lambda \|u\|_t}{\|\tilde{r}\|_t} & \text{if } \|\tilde{r}\|_t \neq 0 \\ \infty & \text{if } \|\tilde{r}\|_t = 0 \text{ and } \|\tilde{e}\|_t + \lambda \|u\|_t \neq 0 \\ 0 & \text{if } \|\tilde{e}\|_t + \lambda \|u\|_t = 0 \end{cases} \quad (1)$$

were λ is a parameter (>0) and the norm is $\|x\|_t = \sqrt{\sum_0^t \rho^\tau x^T(\tau)x(\tau)}$; ρ is a used a an exponential forgetting factor (< 1).

At every moment, the controller performance is evaluated for all C_i according to the procedure (see also Fig. 2):

1. The plant is to be under control of a stabilizing controller, even if its performance is poor.
2. From past input/output data compute a fictitious set-point signal $\tilde{r}_i(t) = \tilde{r}_i(C_i, u(\tau)|_{\tau \leq t}, y(\tau)|_{\tau \leq t})$ for each controller C_i . This corresponds to the set-point signal for which, taking into account $y(t)$, the controller would have produced the actual control action $u(t)$.

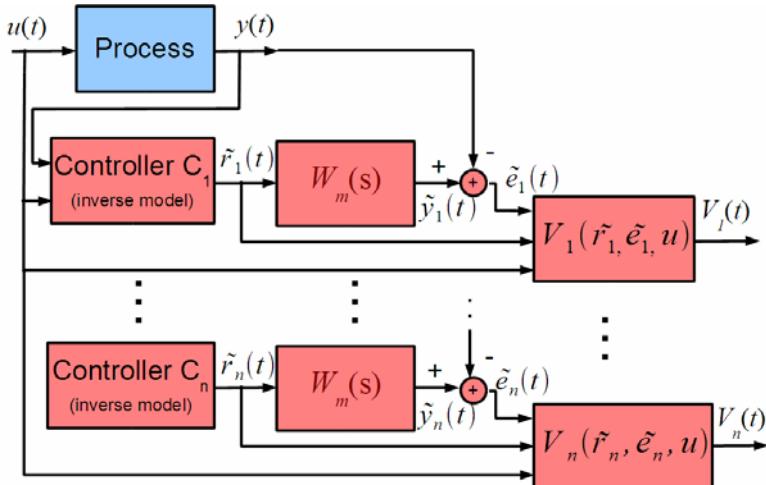


Fig. 2. Controller performance evaluation under unfalsified control framework

3. For each controller compute the fictitious output signal $\tilde{y}(t)$, corresponding to the output of the reference model W_m , when the fictitious set-point signal $\tilde{r}_i(t)$ is used.
4. For each controller compute the fictitious error $\tilde{e}_i(t) = \tilde{y}_i(t) - y_i(t)$.
5. For each controller evaluate the performance function $V(\tilde{r}_i, u, \tilde{e}_i, t)$.
6. From controller C_i performance $V(\tilde{r}_i, u, \tilde{e}_i, t)$ together with a performance threshold γ , the controller is said to be falsified by the available data at time t , if $V(\tilde{r}_i, u, \tilde{e}_i, t) > \gamma$.

3.2 Switch Limiting Strategy

Only unfalsified controllers are candidates to control the process, which means that each individual controller yields a stable closed loop system. The basic control selection approach is to choose the controller with the least performance index. However, this may raise stability problems, resulting from fast switching of the active controller [9, 7]. Thus some switch limiting strategy is required. The most common solution is

the use of a dwell-time. However, two other switching policies are frequently used with unfalsified control switching schemes:

- Once controller C_i is chosen, stick to this controller even though it may not be the best. Controller switching occurs only when $V(\tilde{r}_i, u, \tilde{e}_i, t)$ rises above a threshold, at which time the controller with the least performance index ought to be chosen.
- Another strategy is to define a switching offset. In this case when controller C_i is in use, no switching takes place as long as no other controller performance index lies below $V(\tilde{r}_i, u, \tilde{e}_i, t) - \gamma_s$. This is the policy used in the results from this paper.

3.3 Pole Placement Control Design

A relevant feature under the unfalsified control approach is independence from the controller algorithm. As long as the process is working in stable close loop system, it is possible to evaluate each individual controller performance, even though it is not the actual active controller. This allows the used of any causally-left-invertible linear time invariant controllers.

For pole placement design a set of state space models is required (2), representing the process dynamics over the relevant range of operating conditions. This may be obtained using standard identification methods. The spread of models over the operating range is not critical, as long as it adequately captures the full range.

$$\begin{aligned} \mathbf{x}_i(k) &= \mathbf{A}_i \mathbf{x}_i(k-1) + \mathbf{B}_i u(k-1) & i=1, \dots, n \\ y(k) &= \mathbf{C}_i \mathbf{x}_i(k) + D_i u(k) \end{aligned} \quad (2)$$

For each model a controller is design. The controller structure is

$$u_i(k) = -\mathbf{K}_i \hat{\mathbf{x}}_i(k) + N_i r(k) \quad i=1, \dots, n \quad (3)$$

where $\hat{\mathbf{x}}_i(k)$ is the output from a state estimator, $r(k)$ is the set-point, N_i is a parameter and \mathbf{K}_i is a parameter vector. Each controller is designed so that the closed loop system will behave according to some specified dynamics (the same for all the model/controller).

3.4 Frequency Optimization Based Design

With low noise processes, by using experimental frequency response to characterize the dynamic behavior over the selected operating points it is possible to obtain models (non-parametric) that are closer to the process true behavior. From such models controllers may be designed by classical frequency based methods (Nyquist plot, lead-lag compensation, etc.), or by optimization over the frequency based algorithms. An interesting methodology for this purpose is presented in [14], where a two-stage optimization scheme is used.

4 Implementation and Test

Unfalsified multicontroller supervisors were implemented and tested, using both of the proposed structure and design approaches. This section addresses some implementation issues and experimental results. Tests are made over a lab-scale heat/ventilation experiment (Fig. 3). Three operation regimes are defined according to fan speed: low, medium and high speed. Fig. 3. also shows the frequency response corresponding to each of this regimes.

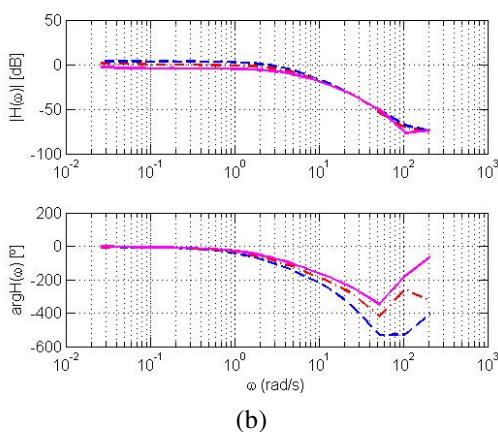
Fig. 3.b shows that at low frequencies the process gain decreases with the fan speed. The process bandwidth increases with the speed. The phase plot shows that the process presents some transport time delay.

4.1 Unfalsified Pole Placement Switched Controller

For each of the selected operation regime a linear second order state space model (2) is identified. Fig. 4 shows a comparison between the process experimental frequency



(a)



(b)

Fig. 3. a) Lab-scale heat/ventilation experiment used for tests. b) Process frequency response for low speed (---blue), medium speed (---red) and high speed (—magenta).

response and that of the identified model. The models gain present a small deviation for low frequencies, but in the overall captures the plant behavior for values above -20dB.

The models phase captures the systems behavior at lower frequencies. However the process transport delay is not captured by the model leading to an increasing difference at higher frequencies. Nevertheless, the model is found to capture the fundamental of the process behavior over the relevant dynamic range.

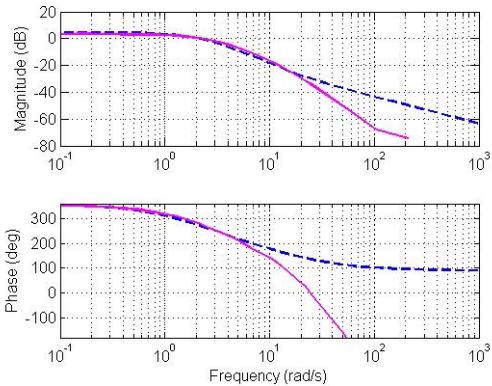


Fig. 4. Process frequency response for low speed (—magenta) and the corresponding state space model frequency response (---blue)

Pole placement controllers (3) are designed according to the closed loop reference model specifications. These are defined in terms of its step response: i) 3% overshoot; ii) settling time of 1 second; iii) zero steady state tracking error. This leads to a reference model represented by the transfer function

$$Y(s) = \frac{16}{s^2 + 6s + 16} R(s) \quad (4)$$

Experimental tests show that the proposed controller structure (3), combined with the identified models, is able to fulfill specifications i) and ii), but yields a significant steady state error. This results from the low frequency modeling error. Thus the control structure from Fig. 1 is adapted to include integral control action (see Fig. 5).

As under unfalsified all the controllers share the same reference model, a shared integral action may be used. If the individual controllers were tuned for different specification sets, it would be necessary to tune a separate control action for each controller.

Fig. 6 illustrates the use of the proposed multicontroller structure applied to the lab-experiment. The test starts at medium fan speed, close to 55 seconds the speed changes to low, and at 105 seconds it changes to high.

The process output follows closely the set-point, with no relevant overshoot and with fast set-point transitions. Operating conditions changes cause disturbances on the output signal that are rapidly recovered. The second fan speed change causes a large spike on the output signal, but it must be stressed that it corresponds to a change from the lower to the higher fan speed. This causes a very large modification on the process gain.

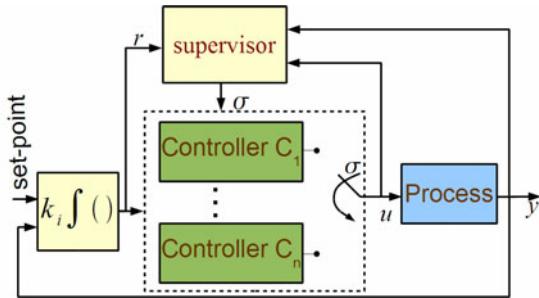


Fig. 5. Supervisory control structure with shared integral action

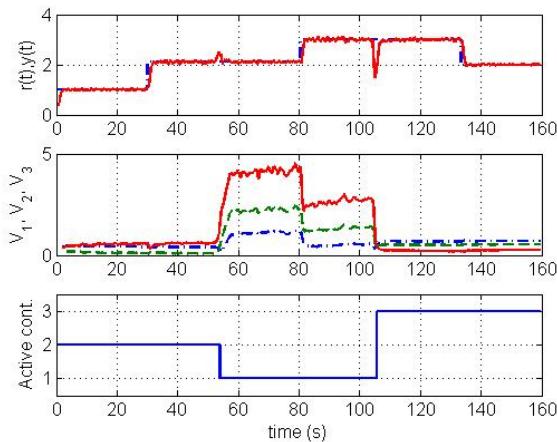


Fig. 6. Supervisory control test over the experimental setup. $r(t)$ - set-point; $y(t)$ – output; V_1 , V_2 , V_3 – performance indexes. Active action: 1 – low speed; 2 – medium speed; 3 – high speed.

Observing the performance indexes (V_1 , V_2 , V_3) it is apparent that the correct controller is selected in all the operating conditions. Operating conditions changes are rapidly detected.

5 Conclusions and Further Work

The switched multicontroller structure described shows to present good performance and fast adaptation to modifications on the operating conditions.

An important feature is that no previous knowledge on the plant dynamics is required to implement the unfalsified control-switching scheme. As the performance evaluation algorithm does not require a specific controller structure, it can be used with a broad range of controllers, and its possible to combine controllers of different types into a single switched multicontroller.

Once the control-switching scheme requires no process model, an interesting development is to design the controllers without using any process parametric models. As mentioned in the paper, this may be achieved through the use of experimental frequency response to characterize the dynamic behavior over the operating range.

References

1. Safonov, M.G., Tsao, T.-C.: The Unfalsified Control Concept and Learning. *IEEE Trans. Aut. Cont.* 42, 843–847 (1997)
2. Safonov, M.G., Cabral, F.B.: Fitting controllers to data. *Syst. & Cont. Letters* 43, 299–308 (2001)
3. Palma, L.B., Coito, F.J., Neves-Silva, R.A.: Robust Fault Diagnosis Approach using Analytical and Knowledge-Based Techniques Applied to a Water Tank System. *Int. J. Eng. Int. Syst. for Elect. Eng. and Comm.* 13, 237–244 (2005)
4. Wang, R., Safonov, M.G.: Stability of Unfalsified Adaptive Control. In: 2005 American Control Conference, pp. 3162–3167 (2005)
5. Hespanha, J.P.: Tutorial on supervisory control. Technical report, Dept. of Electrical and Computer Eng., University of California, Santa Barbara (2001)
6. Morse, A.S.: Supervisory control of families of linear set-point controllers—part 1: exact matching. *IEEE Trans. on Automat. Contr.* 41, 1413–1431 (1996)
7. Morse, A.S.: Supervisory control of families of linear set-point controllers—part 2: robustness. *IEEE Trans. on Automat. Contr.* 42, 1500–1515 (1997)
8. Narendra, K.S., Balakrishnan, J.: Adaptive control using multiple models. *IEEE Trans. on Automat. Contr.* 42, 171–187 (1997)
9. Hespanha, J.P., Morse, A.S.: Certainty equivalence implies detectability. *Syst. & Contr. Lett.* 36, 1–13 (1999)
10. Kosut, R., Lau, M., Boyd, S.: Set-membership identification of systems with parametric and nonparametric uncertainty. *IEEE Trans. on Automat. Contr.* 37, 929–941 (1992)
11. Kosut, R.: Uncertainty model unfalsification: A system identification paradigm compatible with robust control design. In: Proc. of the 34th Conf. on Decision and Contr., pp. 3492–3497 (1995)
12. van Helvoort, J., de Jager, B., Steinbuch, M.: Data-driven multivariable controller design using Ellipsoidal Unfalsified Control. *Syst. & Cont. Letters* 57, 759–762 (2008)
13. Liberzon, D., Morse, A.S.: Basic problems in stability and design of switched systems. *IEEE Contr. Syst. Mag.* 19, 59–70 (1999)
14. Coito, F.J., Ortigueira, M.D.: Fractional Controller Design Trough Multi-Objective Optimization. In: 8th Portuguese Conf. on Aut. Cont. – CONTROLO 2008 (2008)

Design, Test and Experimental Validation of a VR Treadmill Walking Compensation Device

Adrian Stavar¹, Laura Madalina Dascalu¹, and Doru Talaba²

^{1,2} Transilvania University of Brasov, Product Design and Robotics Department,
Bulevardul Eroilor, Nr. 29, Brasov, Romania
{adrian.stavar,madalina.dascalu,talaba}@unitbv.ro

Abstract. Virtual Reality has known exponential development in the recent years and presents important research vectors for the years to come. One major issue, unresolved yet, remains the totally immersion feeling in Virtual Environment (VE) and one of it's most important aspects still to be researched is locomotion in VE. In this direction the authors propose a bidirectional active treadmill based device adapted from an ordinary unidirectional treadmill. The device is used for researching a way to walk smoothly and in an undisturbed fashion on a very limited surface area. For controlling the treadmill's speed and direction the authors propose two simple algorithms. The implementation of the algorithms, experimental setup, tests and results are presented as well. The author's discussion including a critical perspective about the results is also reported. The final part concludes this paper with the authors' perspective and future vectors to be implemented and researched still.

Keywords: Virtual Reality, Immersion, Walking Compensation Device, Control Algorithms.

1 Introduction

Virtual Reality (VR) is one of the newest technological domains that had a flourishing development in the recent years. Many of the VR research directions are rapidly growing and one of the most important is represented by locomotion interfaces.

In VR, locomotion is simulated using several different types of devices, many of them evolved from physical exercise systems. There are several significant aspects that a complete locomotion device must include. One such aspect is the immersion sensation. Together with visual, acoustical and olfactory systems, a walking interface needs to create for the user a total immersion feeling.

A locomotion device used for VR purposes is a system that must provide to the user a manner that allows natural walking in the real environment and precise representation of it in the virtual environment, ensuring an adequate immersion sensation. Influences from the walking system, tracking system and other additional devices have to minimally affect the user's locomotion. Another imperative aspect of a walking simulator is safety which must be set at high standards.

In the simplest VR walking scenario, a mouse, keyboard or joystick controls displacement [1]. More complex scenarios involve tracking the user's position using

cameras attached to a Head Mounted Display (HMD). Different approaches use a Cave Automatic Virtual Environment (CAVE) system comprised of screens used for image projection and magnetic sensors for motion detection.

Several other researchers have been conducting studies on treadmill based devices for use in VR. The first locomotion simulations were done using passive treadmills. In one such case [2] in order to move the rolling band the user uses his strength to push the band, sustained by a pair of auxiliary handlers. Because of user's effort to actuate the rolling band, the locomotion is considered to be unnatural. Subsequently, passive treadmill devices were replaced by active ones [3], [4], [5], [6], [7], [8], [9], [10]. The major advantage of active treadmill locomotion systems is given by the natural and realistic walking sensation felt by the user.

In this paper, we focus on presenting our treadmill based system along with two algorithms used for controlling its speed and sense, adapted to the user's walking style. The user's walking parameters are measured using an electromagnetic Flock of Birds tracker. Real-time capture and processing of these signals is a key feature for the proposed device.

For providing answers to our research questions, tests including the proposed algorithms were developed on eight adult persons. Test method, a short questionnaire, the obtained results and a critical discussion are presented as well.

The final part concludes this paper with the authors' perspectives and future directions still to be researched.

2 Contribution to Sustainability

Human locomotion inside Virtual Environments is an important concept still to be integrated and supported by novel technologies, promising to expand into most of today's domains, like medical-rehabilitation, engineering-architecture, urban planning and also into the concept of tomorrow's Future Internet – the Internet of Services.

The "bridge" that helps achieve virtual walking is represented by locomotion interfaces. By canceling the user's displacement while walking on a limited surface, an infinite new virtual space is available to use and explore. Our small dimensions system manages to reduce the main disadvantage of most active treadmill interfaces - the need for a large walking area [4], [7], [9] – and ensure proper integration in a full immersive CAVE system. To expand functionality, the two control algorithms we propose, properly manage to assist the system to support a natural user's walking sensation. The small dimensions of our proposed bidirectional treadmill system along with its developed control system make the device ideal for future immersion needs.

The goal of the presented system, algorithms and tests is to obtain preliminary data for a complete omni-directional system that we intend to develop.

The proposed system will help future development of already existing domains by opening a path to new ideas and trends for self-sustained and advanced supportive environments. We anticipate that the results presented in this paper will also contribute to VR locomotion interfaces evolution, which represent in our view a small step for sustaining future technologies' development.

3 Walking Compensation Device

3.1 The Compensation System

In this section we present our approach for a walking compensation device, based on a classical jogging treadmill adapted for VR purposes.

The main concept of unlimited walking on a treadmill interface is based on permanently controlling the user position by trying to keep him at a reference position (“dead zone” [4]) set in the middle of the belt’s band. The system always tries to keep the user in this area or very close to it were his movements and locomotion are safe and where the controller applies zero speed to the belt. Any deviation of the user’s centre of mass outside the reference zone is corrected by the control system.

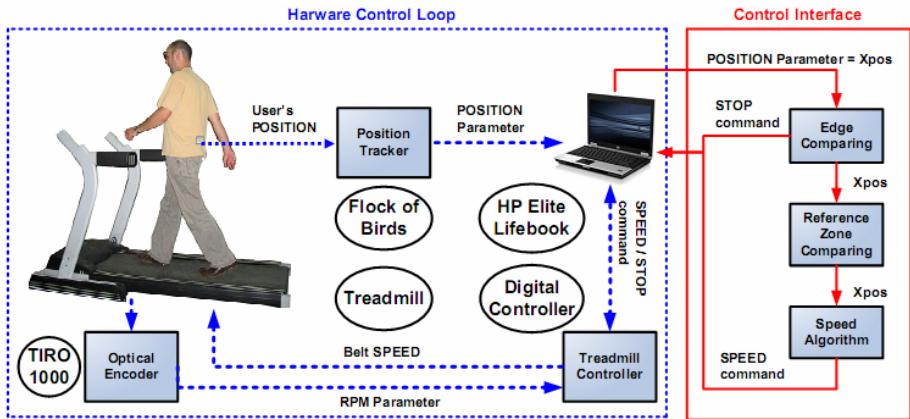


Fig. 1. System Architecture including a *Control Architecture* with the *Hardware Control Loop* (dotted line) and the *Control Interface* (solid line). User’s position parameter (on X axis) during walking is captured with a Flock of Birds (FOB) tracker and sent to *Control Interface*. According to the user’s position relative to reference zone (“dead zone”) and edges of the treadmill’s belt and following the speed control algorithm, the *Control Interface* sends a specific value to the *Treadmill Controller*. This sets the appropriate belt speed. The optical encoder closes the control loop by sending the treadmill’s belt RPM signals to the controller.

Our system was implemented by adapting a classical unidirectional treadmill of 150 cm (L) by 50 cm (W) surface area into a bidirectional VR locomotion device (in Fig. 1, left). From the initial structure only some basic elements were reused, like treadmill structure, belt and DC motor.

The general system architecture (Fig. 1) of the adapted design includes the following subsystems:

- a) HP Elite Lifebook 6930p Laptop, Core2 Duo, 2.40 GHz and 2048 MB RAM
- b) Forward and reverse digital controller – controls precisely the motor’s speed and position and it communicates with the computer through a serial interface (RS232).
- c) Flock of Bird (FOB) magnetic motion tracker system - simultaneously uses 2 tracking sensors to detect the user’s position and head orientation.
- d) Optical incremental quadrature encoder module kit and a 1000 PPR encoder.

The software application includes serial transmission configuration, Flock of Bird data transmission and position reception, parameter recording, position data and control parameters display windows. The two control algorithms were included in the control interface.

The system's safety measures include limits imposed by the algorithms to avoid the case of a user falling off the platform, kill button for stopping the system in dangerous situations and side handlers on the treadmill for stability.

3.2 Control Algorithms

In order for an algorithm to be adequate for walking compensation it has to be able to adaptively control the system by applying an appropriate belt speed opposite to the user's walking speed and sense. Its main role is to minimize the possibility of getting close to the treadmill's edges. In order to test our implemented system we proposed two compensation algorithms, tested in parallel for a better comparison. Our research goal was to determine if we can obtain walking compensation in a limited space area design and what is the proper method to apply.

The first algorithm is based on the idea that when the user starts walking outside the "dead zone" the system has to remotely affect his motion by minimizing inertia and friction forces. The system responds by permanently detecting and comparing the user's position on the walking belt with a set of predefined zones used for transmitting to the treadmill's controller the correct belt speed values. For each forward-backward walking direction, the area is divided into eight speed regions, the first five of low speed increment and the last three of fast speed increment. Dividing the walking area into slow speed system response in the first half and fast speed system response in the last half was based on the idea that the user's balance is affected more in the starting moment and in the first few steps than when the speed is medium and the walking cycle becomes regular.

The second algorithm is based on the idea that the system response has to be smooth not only when the user starts walking but on the entire walking surface. The walking area was divided into 95 equal speed regions for each walking sense. The speed was increased slowly when the user position covered these regions. This slower-linear system reaction had a better influence on the walking compensation than in the first algorithm case.

The pseudo code of the 1st algorithm (Xpos is the detected user position along the X axis)

```

IF: (+Xpos > +edge) OR (-Xpos < -edge)
    STOP Belt
IF: (Xpos is in "Dead Zone")
    STOP Belt
IF: (Xpos is in 1st walking region)
    BELT_SPEED1 = coefficient0
IF: (Xpos is in 2nd walking region)
    BELT_SPEED2 = BELT_SPEED1 + coefficient1
.....

```

```

IF: (Xpos is in 6th walking region)
    BELT_SPEED6 = BELT_SPEED5 + coefficient1
IF: (Xpos is in 7th walking region)
    BELT_SPEED7 = coefficient2
IF: (Xpos is in 8th walking region)
    BELT_SPEED8 = coefficient3
IF: (Xpos is in 9th walking region)
    BELT_SPEED9 = coefficient4

```

The pseudo code of the 2nd algorithm (Xpos is the detected user position along the X axis)

```

Step = f (Area Length, "Dead Zone" Length)
IF: Xpos > 0
    SET Walking Sense
IF: (Xpos < edge) THEN
    Distance = Xpos - "Dead Zone"
    Step Number = f (Step, Distance)
    SPEED = coefficient + Step Number
ELSE: STOP Belt

```

3.3 Testing Method

In order to validate the compensation system's functionality and to determine which of the control methods is the adequate solution for further development we conducted a testing session that included recording and comparison between walking parameters and a questionnaire based evaluation. The tests were performed without including any VR image projection system.

Eight young healthy adults participated in the test (three women and five men) with ages between 24 \div 26 years (25.375 ± 0.74), heights between 164 \div 182 cm (176.125 ± 6.49) and weights between 52 \div 82 kg (67 ± 10.32).

In order to measure and record the walking parameters, a FOB receiver sensor was attached on the user's lower back very close to the center of mass.

Three of the users were somewhat familiar with walking on the treadmill. The remaining five were not familiar at all. All of them were allowed to adapt with the system for 10 minutes, doing a free walk session. After the accommodation session 5 minutes resting was allowed before tests began. The two algorithms were tested sequentially without any pause in between. The test included two forward (0 – 40 s, 60 – 100 s) and two backward (40 – 60 s, 100 – 120 s) free walking sessions. The application recorded the successive FOB receiver positions, speeds and accelerations during free walking and also the belt's speeds at every 125 ms.

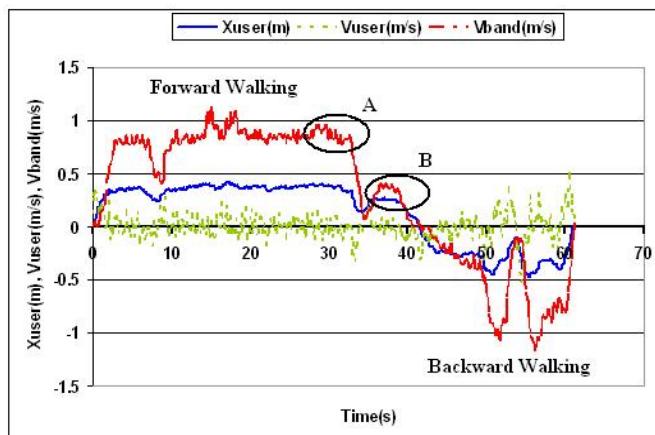
At the end of the testing session a short questionnaire (five questions) was conducted in order to obtain some prompt answers about the walking naturalness and general feelings about the walking compensation system.

3.4 Test Results

The average values of the users' speed, acceleration and band speed both for backward and forward algorithms are represented in Table 1. Standard errors and standard deviations are represented as well for each case.

Table 1. Comparative tests locomotion parameters

		Users' Walking Speed	Users' Walking Acceleration	Treadmill's Belt Speed
1st Algorithm	Mean	0.093 (m/s)	0.521 (m/s ²)	0.681 (m/s)
	Std. Error	0.008	0.045	0.009
	Std. Deviation	0.017	0.091	0.244
1st Algorithm	Mean	-0.097 (m/s)	-0.551 (m/s ²)	-0.511 (m/s)
	Std. Error	0.014	0.086	0.016
	Std. Deviation	0.028	0.172	0.259
2nd Algorithm	Mean	0.104 (m/s)	0.529 (m/s ²)	0.731 (m/s)
	Std. Error	0.010	0.052	0.008
	Std. Deviation	0.020	0.105	0.216
2nd Algorithm	Mean	-0.110 (m/s)	-0.583 (m/s ²)	-0.568 (m/s)
	Std. Error	0.018	0.086	0.011
	Std. Deviation	0.036	0.172	0.211

**Fig. 2.** A user's walking recording: (A) faster compensation, (B) slower compensation, user's displacement (solid line), user's speed (dotted line) and band's speed (dash-dot line)

In Fig. 2 the walking pattern of the 1st user, including his speed and the treadmill's belt speed, is represented.

From the questionnaire responses it resulted that all users were healthy and did not suffer from any mobility disorder. The responses regarding differences between the two walking sessions were: 50 % small, 37.5 % average and 12.5 % no differences. The first algorithm has been seen as reacting a little slower than the second and imposing a slightly insecure locomotion. For the second algorithm it was thought faster than the first in speed and reaction, more adapted and walking was more natural and comfortable. Some general issues regarding both algorithms were expressed. In both cases a small walking imbalance appears when changing walking sense (from forward to backward walking). Also the 5 cm "dead zone" was reported as being too narrow. 62.5 % did not lose their balance, 25 % lost their balance a few times in the 1st session/algorithm, and 12.5 % lost balance once in the 2nd algorithm. Considering

walking naturalness, 75 % suggested that the second algorithm is more adapted, 12.5 % stated the first algorithm and 12.5 % said that there is no difference between them.

3.5 Test Discussion

The results noted in Table 1 show that all the average parameters, including users' speed and acceleration as well as band's speed, are higher for the second algorithm case. Thus, we can state that the moving belt response counteracting the user's walking advance is better in the second algorithm. In the first algorithm walking on the first one half of the active area is softly compensated and the speed increases roughly on the last half. Because belt speed is lower in the first part the user tends to increase his speed more in order to force the system to give a better response. The system reaction is better in the second algorithm because it reacts linear and faster to the user's displacement.

Analyzing the questionnaire answers we can assert that there are no major differences between the two algorithms. Most of the users had comfortable walking sessions without losing their balance, although most of them indicated that the second algorithm is compensating better, giving them a more natural feeling.

Even though we have implemented two different walking compensation approaches with fair results, we can sustain that a proper compensation is also dependent of system hardware and especially of data transmission between the controlling application and system controller. A slower data transmission is a major drawback that can be overcome only by a fast and precise algorithm.

We can conclude that an ideal algorithm has to balance walking speed with limiting inertia and unaffected walking. Ideally the band's speed and the user's walking speed should be almost alike. This means that the system must react very fast to any user step outside the "dead zone" area. But this affects user locomotion start. Thereby it should be a compromise between system reaction and walking disruption.

4 Conclusion and Future Work

Treadmill based systems for simulating locomotion remains one of the major research issues for the Virtual Reality. Creating a totally immersion feeling for the user during traveling in Virtual Environments is an important goal to be researched still.

The question, if a limited area treadmill can provide a proper walking compensation for a small to moderate unaffected and natural walking, arises. In this paper we focus on providing some answers to this question.

A classic treadmill adaptation for VR purposes and the resulting system hardware and software components are presented. Two control algorithms needed for compensating the user's locomotion are also described.

A test and questionnaire on eight young adult persons was conducted to give an opinion about the proper method of walking compensation including the developed algorithms. Test results, including a higher compensating treadmill's band speed and a higher users' walking speed, as well as users' answers of 75 % indicating that the 2nd implemented algorithm induces a more natural walking sensation, led to the conclusion that the 2nd method was more appropriate for walking compensation.

As a future development direction we intend to apply a fuzzy logic control algorithm in order to improve the system's performance. Also we intend to implement a rotational platform integrated in the compensation system to achieve omni-directional walking. Our intention is to research and develop a complete virtual locomotion system with application in actual and future areas of interest.

Acknowledgments. This paper is supported by the Sectoral Operational Programme Human Resources Development (SOP HDR), financed from the European Social Fund and by the Romanian Government under the contract number POS-DRU/6/1.5/S/6 for the authors (1) and by the research project IREAL contract no. 97/2007, id: 132, funded by the Romanian Council for Research CNCSIS for the authors (2).

References

1. Kim, G.J.: Designing VR Systems, The Structured Approach, pp. 3–5. Springer, London (2005)
2. Patel, K.K., Vij, S.K.: Unconstrained Walking Plane to Virtual Environment for Spatial Learning by Visually Impaired. *Ubiquitous Computing And Communication Journal* (2010)
3. Darken, R.P., Cockayne, W.R.: The Omni-Directional Treadmill: A Locomotion Device for Virtual Worlds. In: Proc. UIST 1997, pp. 213–221 (1997)
4. Iwata, H.: Walking About Virtual Environments on an Infinite Floor. In: Proc. of the IEEE Virtual Reality (1999)
5. Iwata, H.: Art and Technology in Interface Devices. In: Proc. of the ACM Symposium on Virtual Reality Software and Technology, pp. 1–7 (2005)
6. Christensen, R., Hollerbach, J.M., Xu, Y., Meek, S.: Inertial Force Feedback for the Treadport Locomotion Interface. In: Presence: Teleoperators and Virtual Environments, vol. 9, pp. 11–14 (2000)
7. Hollerbach, J.M.: Locomotion Interfaces. In: Stanney, K. (ed.) *Handbook of Virtual Environments Technology*, pp. 239–254 (2002)
8. Noma, H., Miyasato, T.: A New Approach for Canceling Turning Motion in the Locomotion Interface, ATLAS. In: Proc. of AME-DSC, vol. 67, pp. 405–406 (1999)
9. Giordano, P.R., Souman, J.L., Mattone, R., De Luca, A., Ernst, M.O., Bulthoff, H.H.: The CyberWalk Platform: Humna-Machine Interaction Enabling Unconstrained Walking through VR. In: First Workshop for Young Researchers on Human-Friendly Robotics (2008)
10. De Luca, A., Mattone, R., Giordano, P.R., Bulthoff, H.H.: Control Design and Experimental Evaluation of the 2D *CyberWalk* Platform. In: IEEE/RSJ IROS 2009, pp. 5051–5058 (2009)

Design, Manufacturing and Tests of an Implantable Centrifugal Blood Pump

Eduardo Bock^{1,2,3}, Pedro Antunes^{1,4}, Beatriz Uebelhart^{1,5}, Tarcísio Leão³, Jeison Fonseca^{1,6}, André Cavalheiro⁴, Diolino Santos Filho⁴, José Roberto Cardoso⁴, Bruno Utijama¹, Juliana Leme¹, Cibele Silva¹, Aron Andrade¹, and Celso Arruda²

¹ Institute Dante Pazzanese of Cardiology, IDPC

² Unicamp, Campinas State University

³ Federal Institute of Technology, IFSP

⁴ Escola Politecnica of São Paulo University, EPUSP

⁵ Faculty of Technology, FATEC

⁶ Technological Institute of Aeronautics, ITA, Brazil

eduardo_bock@yahoo.com.br

Abstract. An implantable centrifugal blood pump was developed for long-term ventricular assistance in cardiac patients. In vitro tests were performed, as wear evaluation, performance tests and hemolysis tests in human blood. Numerical computational simulations were performed during design process in order to predict its best geometry. Wear evaluations helped to select the best materials for double pivot bearing system proposed to achieve longer durability. Performance tests pointed the best impeller geometry. The implantable centrifugal blood pump was compared with other blood pumps founded in literature. The proposed implantable centrifugal blood pump showed the best performance. But, its results showed a strong descendant curve in high flow. Other prototype was manufactured with a different inlet port angle to overcome this problem. The normalized index of hemolysis (NIH) measured 0.0054 mg/100L that can be considered excellent since it is close to the minimum found in literature (between 0.004 g / 100L e 0.02 g / 100L). The authors' expectation is that this pump will become a promising Technological Innovation for Sustainability.

Keywords: Artificial Organs, Blood Pumps, Ventricular Assist Devices.

1 Introduction

A novel Implantable Centrifugal Blood Pump has been developed for long term circulatory assistance with a unique impeller design concept [1]. This feature was called dual impeller because it allies a spiral-shaped cone [2] with vanes to improve blood flow characteristics around the top inflow area to avoid blood clot due to stagnant flow (Fig. 1).

A series of previous studies demonstrated significant advantages from spiral shaped impeller design, providing axial characteristics to the flow in Left Ventricle Assist Devices [1]. The axial force component can avoid stagnant flow formation. Therefore, this principle can help to avoid thrombus related with blood stagnation [3].

This work presents results from Design, Manufacturing and Tests of an Implantable Centrifugal Blood Pump wear evaluation in double pivot bearing system, hydrodynamic performance tests with mock loop circuit, and preliminary normalized hemolysis tests.

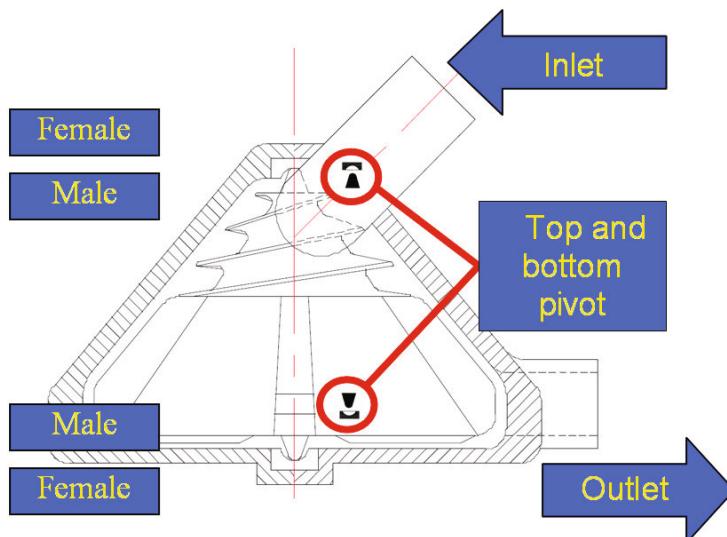


Fig. 1. The new Implantable Centrifugal Blood Pump for cardiac assistance is shown with its double pivot bearing system composed by male and female bearings

The double pivot bearing system has been used and studied in the last decade showing simplicity and reliability [4]. Recent studies evaluated the particle release from a centrifugal pump with double pivot bearing system composed of alumina ceramic (Al_2O_3) and ultrahigh molecular weight polyethylene (UHMWPE). This pivot bearing system was tested under severe conditions showing very small risk of releasing debris particles to blood [1, 5].

2 Contribution to Sustainability

2.1 The Design Process

The design process was conducted according with current medical, social and financial needs of cardiac patients in Brazil. It is part of a multicenter and international study with objective to offer simple, affordable and reliable devices to developing countries since local health systems, both public and private health care, cannot afford the available technologies.

The authors have expectation that this pump will become a promising Technological Innovation for Sustainability in Brazil. Its simplicity is illustrated in (Fig. 2) showing two different options for motors and pump.

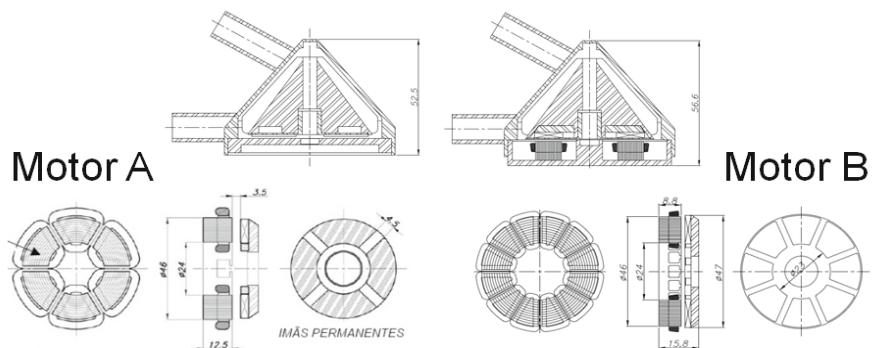


Fig. 2. The design process was lead according to current needs of cardiac patients who cannot afford Left Ventricular Assist Device technology. The reduced number of parts is shown in two different options of motor attached with pump's rotor by a magnetic coupling.

2.2 Performance Tests

Performance tests were conducted with several different prototypes, some of them machined in translucent materials and titanium, other made with Selective Laser Sintering (SLS) in Nylon and ABS polymer (Fig. 3 shows some pictures of prototypes).



Fig. 3. Different types of prototypes obtained by SLS or Machined in different types of materials from translucent polymers to titanium

In order to verify the viability of different materials, wear tests were made with several pivot bearings. Mainly, two types of defect were found, abrasive wear and surface fatigue. Previous tests with this system were performed with pumps working in mock loop circuits.

In order to study the wear phenomenon in each component of double pivot bearing, an isolated wear station was assembled to repeat this condition. In vitro performance tests were made in order to characterize hydraulic performance curves for the pump.

These tests can provide important information about the pump's capability and function ability. The generated curves can be used as a tool to predict which rotation is necessary to provide specific pressure and flow [6].

During performance tests, two different types of pump's inlet port were compared, with 45° and 30° of inlet angle. Finally, two isolated normalized tests of hemolysis with human blood were performed producing four values of normalized index of hemolysis (NIH), obtained from variation of plasma free hemoglobin (PFH), measured by a tetramethylbenzidine (TMB) assay method (Catachem, Bridgeport, CT, USA).

3 Materials and Methods

Wear Evaluation in Double Pivot Bearing System. A wear test station was assembled with the purpose of measuring isolated wear rate. It makes it possible to vary the shear stresses in consequence of charge applied to the system. The main idea was to evaluate wear using mass measurements and quantifying the bearing's material debris released in blood during the pump's operation [1].

Adapting a milling machine, the wear test station was assembled with rotation controller, water lubrication, depth controller, and applied charge measurement system. The wear tests were divided into three steps. In the first step, all bearings were weighted in a precision scale with divisions of 0.1 mg. After that, each pair of male and female bearing was tested. Two types of pairs were tested, ceramic–polymeric pairs and ceramic–ceramic pairs. After the tests, the pairs were weighted again to measure the wear loss in mass.

The polymers chosen were nylon, UHMWPE, and Teflon (poly-tetrafluorethylene). The ceramics chosen were zirconium dioxide (ZrO_2), silicon nitride (Si_3N_4), carbon, and alumina (Al_2O_3).

All pieces tested had their contact surfaces polished and roughness controlled to assure an average of 0.1 to 0.5 mm. Each test was performed under the following conditions: room temperature 21°C, 9.8 N of charge applied at 4000 rpm for a total of 40 000 revolutions.

Hydrodynamic Performance Tests with Mock Loop Circuit. A performance test circuit was assembled with one hanging flexible reservoir sac (Flexible Sac, 3M, St. Paul, Minnesota, USA), two pressure probes, one pressure monitor, one ultrasonic flowmeter with 3/8" probe (Transonic Systems, Ithaca, NY, USA), pump and controller, acquisition PCI slot, and a laptop with Labview software (National Instruments, Austin, TX, USA).

The mock loop circuit was set with water solution with 37% of volume filled with glycerin to simulate the blood viscosity and density. The hanging flexible reservoir was filled with 0.4 L and placed 0.5 m above the pump. The pressure gauges were

connected 0.3 m from pump inlet and outlet ports. The ultrasonic flowmeter transducer was located 0.15 m from pump outlet. Total length of each tubing from pump inlet and outlet to reservoir sac was 1.0 m, Fig. 4. The mock circuit was set horizontal to the table for easy adjustment of pump afterload using a screw clamp. Any air was removed from the circuit before the data acquisition [7].

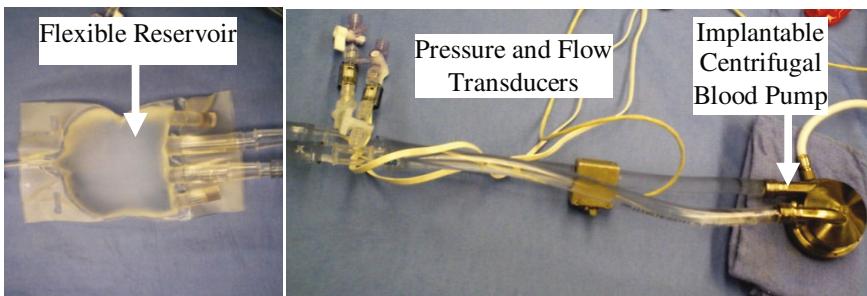


Fig. 4. Simple mock loop circuit assembled for data acquisition during performance tests

Two pumps were tested with different inlet angles, 45° and 30°. The angles were chosen according to similar devices. The pump speed rotations were fixed at 1200, 1400, 1600, 1800, 2000, and 2200 rpm. The software plotted a curve for each rotation, and for each pump. With screw clamp open, the first point collected was the maximum flow for each pump speed. Closing the screw clamp, the pump afterload increases as the flow decreases, and the next points are collected each 0.5 L/min, successively until 0.0 L/min is achieved, when the clamp is totally closed.

The pump speed rotations were fixed at 1200, 1400, 1600, 1800, 2000, and 2200 rpm. The software plotted a curve for each rotation, and for each pump. With screw clamp open, the first point collected was the maximum flow for each pump speed. Closing the screw clamp, the pump afterload increases as the flow decreases, and the next points are collected each 0.5 L/min, successively until 0.0 L/min is achieved, when the clamp is totally closed.

Normalized Index of Hemolysis (NIH) Tests. According to ASTM Standard Practices F1830 and F1841, the hemolysis test was divided into five steps: collection and preparation of blood, mock loop setup, 6 h test, PFH measurement, and NIH calculation.

Blood was drawn from human volunteer using a large bore needle into a 500 mL blood bag with heparin to avoid clot formation during all procedures. No negative pressure exceeded 100 mm Hg and the blood temperature was controlled between 2°C to 8°C.

A test loop was assembled with 2 m of flexible 3/8" polyvinyl chloride tubing, a flexible reservoir with sampling port, an ultrasonic flowmeter with probe, a pressure monitor with probes, a thermistor with thermometer, and the blood pump (similar to performance station shown in Fig. 4). Blood had the hematocrit range between 28% to 32% controlled by hemodilution with phosphate-buffered saline.

Temperature was adjusted to 37°C. The pump rotation was adjusted to provide 5L/min flow rate. A screw clamp was set around flexible tubing at the pump outlet

side to produce the required pressure conditions for left ventricle assist devices (100 mm Hg). In each 6 h test, seven samples were collected T0,T1,T2,T3,T4,T5, and T6, and their respective PFH was measured.

4 Results

Ranking. The results of wear evaluation in different materials tested were sorted by loss of mass, and a ranking list of 10 best results was plotted in Figure 5. Values express the total loss of mass caused by wear in pairs and its total weight, with standard deviation of 0,000022 for measures.

Performance curves. A diagram (Flow vs. Pressure Ahead, as shown in Figure 6) was plotted with each pump curve superimposed in order to better understand the differences between both hydrodynamic performances.

	Male bearing	Female bearing	Wear loss (g)
1st	Alumina	UHMWPE	0.0001
	Silicon nitride	Silicon nitride	0.0001
	Carbon	UHMWPE	0.0001
	Carbon	Nylon	0.0001
5th	Alumina	PTFE	0.0002
	Silicon nitride	Nylon	0.0002
	Carbon	PTFE	0.0002
8th	Alumina	Nylon	0.0003
10th	Zirconium dioxide	UHMWPE	0.0003
	Zirconium dioxide	Nylon	0.0006

UHMWPE, ultrahigh molecular weight polyethylene; PTFE, poly-tetra-fluoroethylene (Teflon).

Fig. 5. Ranking of materials under wear tests

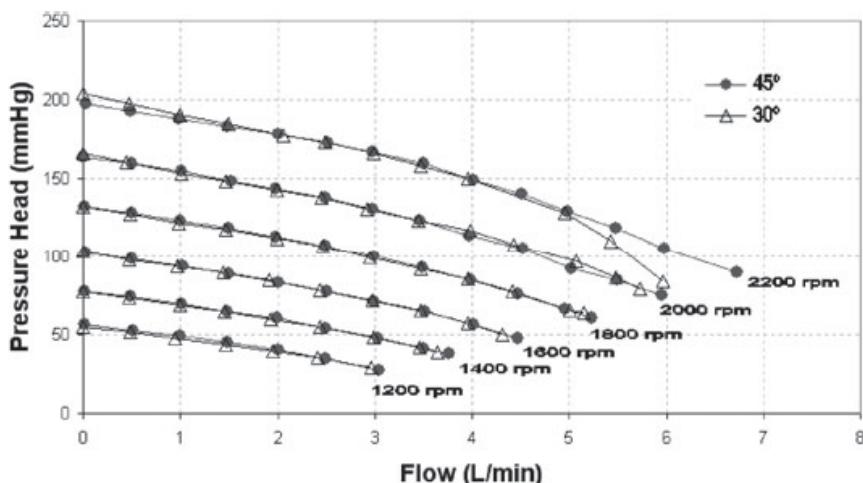


Fig. 6. Performance curves showing hydrodynamic characteristics for both pumps

PFH releasing. PFH value was measured for each sample. With this PFH variation in time T0, T1, T2, T3, T4, T5, and T6 is possible to represent hemoglobin releasing in plasma during the experiment. This progressive alteration in PFH levels in blood is caused by trauma imposed by pump.

5 Discussion

As described in previous studies [1], the bearing composed of alumina with UHMWPE showed the best results in wear evaluation trials with ceramic–polymeric pairs [8]. Four pairs had the minimum mass loss (0.1 mg): alumina with UHMWPE, silicon nitride with silicon nitride, carbon with UHMWPE, and carbon with nylon.

Pivot bearing systems composed only of ceramic are known to have higher vibration during pumping applications instead of shock absorption experienced in ceramic–polymeric pivot bearing systems [4, 7].

The pump with inlet angle of 45° showed best performance results compared with other pumps built with inlet angle of 30°. Slight differences among curves were found in rotations beyond 1800 rpm when the curves from 30° pump decrease pressure ahead versus flow. Inlet port's angle is a problem to deal with when designing centrifugal blood pumps, especially the eccentric inlet port [9, 10].

As described by several authors, NIH for LVAD should be between 0.004 to 0.02 mg/100 L to be considered satisfactory and antitraumatic blood pump [1, 6, 10]. After calculations, the preliminary hemolysis tests showed an NIH value of 0.0054 $\pm 2.46 \times 10^{-3}$ mg/100 L.

6 Conclusions

The pair composed of alumina and UHMWPE was chosen to be the materials of the double pivot bearing system in order to avoid vibration problems.

The dual impeller centrifugal blood pump had proven to be a promising LVAD. The hydraulic characteristics are similar to other reported curves from established, durable, and reliable devices.

The authors' expectation is that this pump will become a promising Technological Innovation for Sustainability.

Acknowledgments

The authors are grateful to Dr. Yukihiko Nosé for enormous collaboration to this project. Also, it is mandatory to thank all colleagues from Baylor College of Medicine (BCM), Federal Institute of Technology (IFSP) as well as CNPq, HCor and Adib Jatene Foundation (FAJ) for partially supporting this research.

References

1. Bock, E., Antunes, P., Andrade, A., et al.: New Centrifugal Blood PumpWith Dual Impeller and Double Pivot Bearing System:Wear Evaluation in Bearing System, Performance Tests, and Preliminary Hemolysis Tests. *Artif. Organs* 32, 329–333 (2008)

2. Andrade, A., Biscegli, J., Dinkhuysen, J., et al.: Characteristics of a blood pump combining the centrifugal and axial pump principles. *Artif. Organs* 20, 605–612 (1996)
3. Yamane, T., Miyamoto, Y., Tajima, K., et al.: A comparative study between flow visualization and computational fluid dynamic analysis for the sun medical centrifugal blood pump. *Artif. Organs* 28, 458–466 (2004)
4. Ohara, Y., Sakuma, I., Makinouchi, K., et al.: Baylor Gyro pump: a completely seal-less centrifugal pump aiming for long-term circulatory support. *Artif. Organs* 17, 599–604 (1993)
5. Takami, Y.: In vitro study to estimate particle release from a centrifugal blood pump. *Artif. Organs* 30, 371–376 (2006)
6. Takami, Y., Andrade, A., Nosé, Y., et al.: Eccentric inlet port of the pivot bearing supported Gyro centrifugal pump. *Artif. Organs* 21, 312–317 (1997)
7. Hansen, E., Bock, E., Nosé, Y., et al.: Miniaturized all-in-one rpm controllable actuator for gyro centrifugal blood pump. In: 52nd ASAIO Annual Conference, Chicago (2006)
8. Takami, Y., Nakazawa, T., Makinouchi, K., et al.: Material of the double pivot bearing system in the Gyro C1E3 centrifugal pump. *Artif. Organs* 21, 143–147 (1997)
9. Andrade, A., Biscegli, J., Souza, J., et al.: Flow visualization studies to improve the spiral pump design. *Artif. Organs* 21, 680–685 (1997)
10. Nosé, Y.: Design and development strategy for the rotary blood pump. *Artif. Organs* 22, 438–446 (1998)

Embedded Intelligent Structures for Energy Management in Vehicles

Ana Pușcaș¹, Marius Carp¹, Paul Borza¹, and Iuliu Szekely²

¹ Transilvania University of Brașov, Electrical Engineering and Computer Science
29 Eroilor Street, 500036, Brașov, Romania

² Acta Universitatis Sapientiae,

1 Sighisoarei Street, 540485, Targu Mureș, România

{ana_maria.puscas,marius.carp}@yahoo.com, paul.borza@unitbv.ro,
gszekely@ms.sapientia.ro

Abstract. The present research is focused on developing embedded intelligent structures for energy management in mobile systems. The research introduces a hybrid structure composed of different energy sources endowed with control systems able to optimize the power flow of the ensemble. In the paper, the architecture of the proposed hybrid system is described. To test the functionality and the advantages of the system, preliminary simulations were made. To characterize the behaviour of the physical developed system, a test bench was implemented.

Keywords: combined energy sources, energy efficiency, control systems, control strategies, electric vehicle.

1 Introduction

Lately, the automotive industry has been focused on reducing fuel consumption, energy consumption and pollutant emissions while increasing the global efficiency of the vehicle and the passengers' comfort. The new facilities for improving the comfort (HVAC system of the vehicle, GPS, night vision camera, parking assistant etc) increase the energy and fuel consumption, thus heating the battery, reducing its lifetime and performances and affecting the environment.

In order to respect environmental issues, the general trend in the automotive industry is to transit from the classic vehicles toward hybrid electric vehicles (HEV) and electric vehicles (EV), non-polluting and economical vehicles [1].

Starting 1997, Toyota and Honda opened the gates in HEV and EV. At present, efforts are being made for the transition to pure EV (battery EV and fuel cell EV). In this sense, both Honda and Toyota companies have developed models certified by the U.S. Environmental Protection Agency (EPA) and California Air Resources Board (CARB) as being Partial Zero Emission Vehicle (PZEV) and ZEV.

The final aim of the present research is to optimize the energy and fuel consumption of the HEV/EV, the lifetime of the storage devices and to increase energy efficiency. As a result, the autonomy and the dynamism of the vehicles have to be increased. In order to fulfil the final aim, the goal of the paper is to study and optimize

the power flow between the electrical and mechanical sides of an EV. By considering a combined energy cell (CEC), DC-DC converter, electrical machine (DC motor/generator - MG) and flywheel that simulates the inertia of the EV, a reduced scale model was developed before a detailed analysis of the components was performed. A CEC represents a hybrid system composed of different storage and generation devices with different time constants (battery, supercapacitor - SC), all embedded in an intelligent structure with computation abilities.

Starting from a theoretical approach, a model of a network of four CEC cells was simulated. Based on the results of the simulations, a first image about the implementation was achieved. In the experimental phase, a test bench for a reduced scale physical model that includes one CEC was designed and implemented. Also, an embedded intelligent structure for energy management (EISEM) was developed. To facilitate the energy management process of an EV, the EISEM integrates energy storage devices and use embedded systems based on microcontrollers (UC) and network of switching devices. Using the prototype, the regenerative braking process of the EV was simulated, the experimental results being emphasized in the paper.

2 Contribution to Sustainability

Nowadays, the increase in energy efficiency represents the main target of any emergent technology. In the present research, an innovative technique that reflects the fusion between energy and information implemented as a dynamic network of CEC cells is proposed. The sustainability targets are related to improving the performances of the three level cellular power supply architecture and of the control strategies (laws) that govern it. For the CEC structure (first level), a simple control strategy was considered and implemented. At the network level (second level), the dynamic changing of the topology was simulated. At the vehicle level (third level), the corollary of the research will be achieved by correlating the transport missions with the energy resources of the vehicle.

Using EISEM devices, an optimal balance between the provided power and the energy density is reached. Also, the volume, the wasted energy and the pollution are reduced, such a system being suitable for Start/Stop systems and regenerative braking processes. Based on the experimental and simulation results, the performances of the EISEM are evaluated and its model will be integrated as power supply in the overall HEV/EV models. Considering the economical aspects, this solution represents a good compromise between performances and the costs obtained without major technological efforts. Thus, the prototype is considered an important step toward improving the sustainability of the energetic solutions.

3 Storage Devices Used in Automotive – State of the Art

The development of the HEV/EV is limited especially because of their power supplies and their weak characteristics. Even in the case of NiMH and Li-ion batteries this limitation persists and affects the reliability and the autonomy of the vehicle [2]. The energy efficiency, lifetime, cyclability, dynamical performances and the starting processes are affected by the functioning regimes of the vehicles [3]. In addition, the high

time constant of the battery decreases the performances of the regenerative braking process thus limiting the technological advance in the energy management systems for HEV/EV.

The actual solutions overpassing the above mentioned issues consist in: (i) increasing the vehicle's voltage from 12 V to 42 V; (ii) oversizing the battery's capacity; (iii) hybridization of the power supply and increasing the dynamicity of the vehicles in urban traffic by including fast release storage devices. The disadvantage of the first solution is related to the high costs of the vehicles' technology translation. The disadvantages of the second solution are related to increasing the costs, the carried weight of the system and the pollution. The third solution represents the actual trend in automotive, an optimal and sustainable compromise [4], [5], [6], [7], [8], [9].

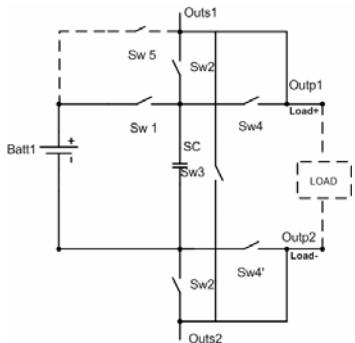
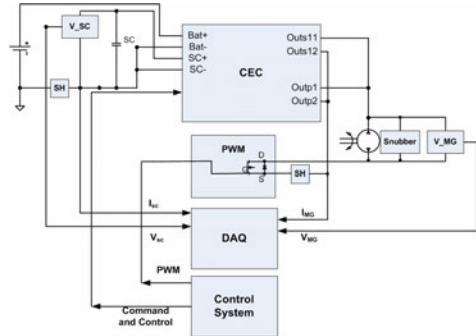
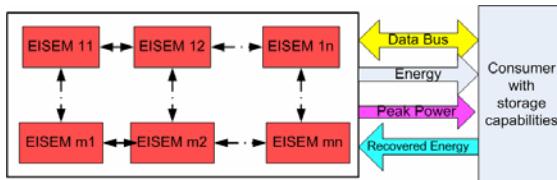
Because there is no device able to ensure high energy and power densities, and increased lifetime at the same time, the automotive industry has focused its attention on developing embedded control solutions for reducing fuel consumption and increasing efficiency [10], [11], [12]. The actual research uses batteries as storage devices but there are also implementations which use SC in order to increase the power density and the lifetime of the ensemble [3], [13], [14], [15].

The present paper describes a hybrid storage and energy device embedded in an intelligent structure. The hybrid system is composed of cells of batteries, SC, sensor networks and intelligent control system able to combine and commute all the storage and generation devices thus optimizing the power flow. The SC has the advantage of being non-polluting and able to provide and smooth high peak current pulses, thus increasing the lifetime of the battery [16]. The main difference between the researched system and the existing ones is the control made at cell level, instead of the control made at device level [17].

4 Hybrid Storage Device – Architecture

By connecting multiple storage devices (batteries, SC, flywheels) characterized by different time constants and performances into embedded intelligent structures, with a view to ensuring long lifetime, high energy density, and high power density of the ensemble, storage and energy device hybridization is obtained (Fig. 1) [18]. For increasing the performances of the ensemble, the hybrid system was endowed with an intelligent control system able to monitor, control and optimize the energy flow of the vehicle, thus improving its performances. The architecture of the hybrid structure, illustrated in Fig. 2, is composed of EISEM connected to a MG controlled through PWM. The EISEM consists of: (i) CEC: hybrid device composed of cells of storage devices and back to back switches able to ensure the bidirectional power flow (Fig. 1); (ii) Data Acquisition System (DAQ): embedded system based on ATMega128 UC used for acquiring data from the voltage, current and revolution sensors; (iii) Control System: embedded system based on UC for controlling and optimizing the power flow transferred through CEC.

Multiple EISEM can be connected in series/parallel to increase the performances of the system (Fig. 3).

**Fig. 1.** CEC cell**Fig. 2.** Architecture of the hybrid structure**Fig. 3.** Network of EISEM

5 Simulations and Results

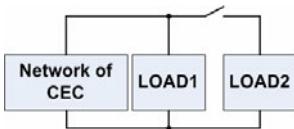
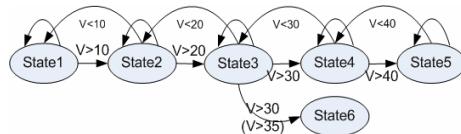
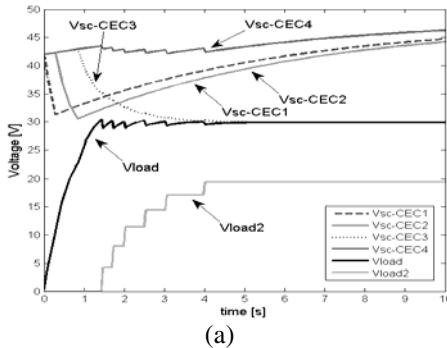
The main advantages of a hybrid storage device are related to: (i) increasing the lifetime of the battery by supplying the high peak pulses from the SC in the starting process and (ii) improving the energy efficiency of the system in the regenerative braking process because of the reduced time constants of the SC [19].

To highlight the performances and the advantages of a network of CEC cells (Fig. 3), the system was modelled and simulated with Matlab/Simulink tool. The architecture of the modelled system is illustrated in Fig. 4. The simulated network was composed of four CEC cells which can be connected in series and/or in parallel in order to ensure the load profile. In the simulated hybrid device composed of four CEC cells, four 400 F/42 V SC and four 77 Ah/42 V lead acid batteries were used. Also, as analogy with the inertial storage from HEV/EV, two capacitive loads ($Load1 = 400$ F, $Load2 = 200$ F) were considered.

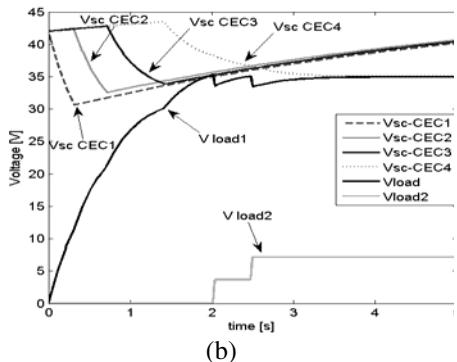
The control strategy (Fig. 5) was implemented to charge the $Load1$ from 0 V to 40 V and to supply $Load2$ from $Load1$. To control the charging process, voltage transducers were used.

The results of the simulations are illustrated in Fig. 6 a), b).

The results of the simulations are in accordance with the logic described by the control algorithm (Fig. 5). Depending on the voltage of the load, the network of cells was commuted. In *State1*, $Load1$ is charged from SC-CEC1 until its voltage reaches 10 V. In *State2*, $Load1$ was charged up to 20 V from the SC-CEC2, in *State3*, $Load1$ was charged up to 30 V from the SC-CEC3. From *State4* started the stabilization

**Fig. 4.** Simulated system**Fig. 5.** Control strategy used in simulations

(a)



(b)

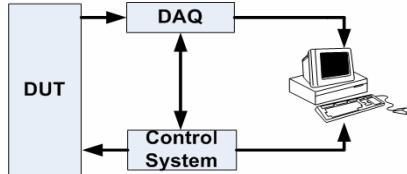
Fig. 6. Results of the simulation

phase; the algorithm jumped in *State5* and *State6* in order to supply *Load2* while maintaining the voltage on *Load1*. Thus, for voltages measured on *Load1* greater than 30 V (Fig. 6 a) and 35 V (Fig. 6 b), *Load2* was supplied from *Load1*. At the same time, the CEC cells were used for maintaining the 30 V (respectively 35 V) level on *Load1*. The simulations proved that a network of CEC hybrid devices endowed with the adequate control strategies can be used for mastering the desired load profile. This embedded system reflects the fusion between the power flow transmitted bidirectionally between generator - load and the necessary information for the control strategies.

6 Experiments and Results

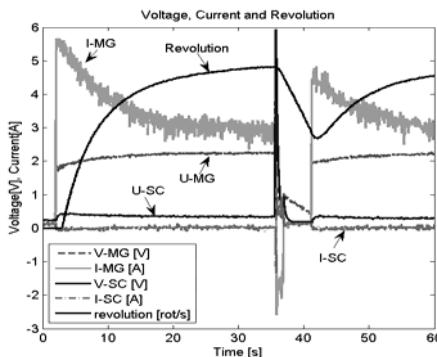
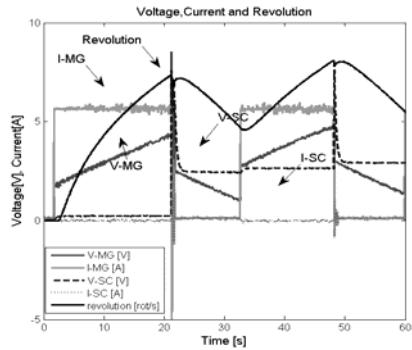
A single EISEM cell was implemented and tested before the physical implementation of the network of cells. In order to experimentally test the functionality of the prototype (EISEM hybrid device) (Fig. 2) and its efficiency, a first test bench was implemented. The test bench (Fig. 7) is composed of: Device Under Test (DUT), DAQ, Control System and laptop.

The DUT module consists of: CEC hybrid system, 120 W MG with 0.27 kg/m^2 inertia, snubber circuit, Hall sensors for monitoring the values of the MG and SC currents, sensors for monitoring the MG and SC voltages and optical sensors for monitoring the revolution of the MG. A 0.11 F small SC was used as generation and storage device in CEC. The battery was simulated with a DC voltage source limited at 2.5 A and 5 A.

**Fig. 7.** Test Bench

The system was implemented to ensure the communication between the modules. Data were acquired with DAQ, were sent to a laptop by using Visual Basic software and UART facilities and were locally stored and processed using Matlab tool. Based on the information received from DAQ, the Control System interprets the data and controls both modules DAQ and CEC in order to optimize the process.

The first experiments were made while supplying the MG at the power values of 12.5 W (Fig. 8) and 40 W (Fig. 9). In these experiments, while braking, the energy is recovered and stored on the SC without using a DC-DC converter.

**Fig. 8.** Experiments at 2.5 V and 5 A**Fig. 9.** Experiments at 8 V and 5 A

If a DC-DC converter and adequate control system are integrated in the tested prototype, the recovered energy can be significantly increased (Fig. 10 and Fig. 11).

In the experiments, the regenerative braking efficiency implemented at reduced scale was tested. The energy recovered and available in the SC is given by Eq. 1:

$$E = (1/2) \cdot C \cdot (U_{\max}^2 - U_{\min}^2). \quad (1)$$

By analogy with the urban traffic, a control strategy was conceived. The strategy controls the modality to provide the energy necessary to move the vehicle by accelerating and to recuperate the energy by decelerating. As it can be seen in Fig. 10 and Fig. 11 the energy recovered in 60 seconds of successive accelerations and breakings, while the MG is supplied at 2.5 V/2.5 A is 0.22 J and while the MG is supplied at 8 V/5 A is 1.48 J. Thus, if supplying at 12 V/5 A the energy recovered in 60 seconds is 8 J. For testing a real implementation of regenerative braking on EV, 400 F/14 V

SC able to recover 39.2 kJ will be used. Thus, the energy efficiency can be significantly increased by recovering the energy usually lost into heat.

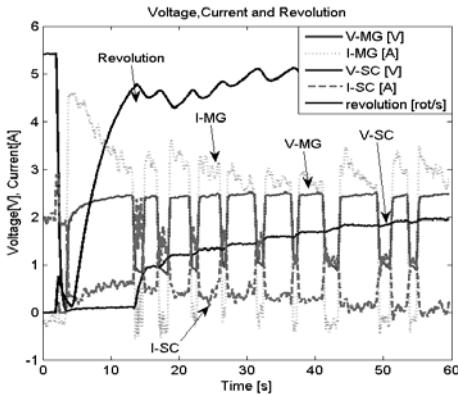


Fig. 10. Experiments at 2.5 V and 5 A

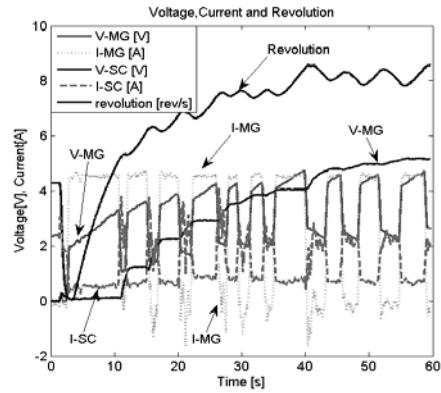


Fig. 11. Experiments at 8 V and 5 A

7 Conclusions and Future Work

This paper briefly presents the advantages of using embedded intelligent structures for increasing the performances of the HEV/EV. The paper demonstrates the viability of a cellular concept used for controlling the power flow from the sources toward loads with storage characteristics, as an alternative to the batteries systems endowed with DC-DC converters. By analogy, this application is similar to the generic control of the HEV/EV where an essential phenomenon consists in recuperating the kinetic energy of the vehicle. The energy efficiency can be increased by replacing the DC-DC converters with such network of CEC cells.

In the present paper, a network of EISEM was introduced and the EISEM hybrid energy storage device was detailed. To observe its behaviour, the EISEM composed of four CEC cells was firstly modelled and simulated. The simulations proved that the EISEM not only can increase the performances of the ensemble, but it also can be endowed with an adequate control system in order to optimize and maintain the load profile.

A first reduced scale physical prototype of the EISEM hybrid configuration was implemented and its architecture was described. The experimental results proved its efficiency while using it in the regenerative breaking and starting processes. Based on the experiments, the corresponding model has to be developed and simulated.

As future work, the network of EISEM cells will be physically implemented. Inside the network, the EISEM cells will have the capability to communicate with a control system and also between them. The control system will have the intelligence to optimally connect the EISEM cells in order to optimize the power flow inside the system. Also, a full scale implementation for EV and HEV has to be done.

Acknowledgments. This paper is supported by the Sectoral Operational Programme Human Resources Development (SOP HRD), financed from the European Social Fund and by the Romanian Government under the contract number POSDRU/6/1.5/S/6.

References

1. Larminie, J., Lowry, J.: Electric Vehicle Technology Explained. John Wiley, Chichester (2003)
2. Settle, F.A. (ed.): Handbook of Instrumental Techniques for Analytical Chemistry. Prentice-Hall, Inc., NJ (1997)
3. Borza, P., Pușcaș, A.M., Székely, I., Nicolae, G.: Energy Management System based on supercapacitors used for starting of internal combustion engines of LDH1250 locomotives and charging their batteries. In: SIITME 2008, Predeal, România, pp. 227–231 (2008)
4. Karden, E., Ploumen, S., Fricke, B., Miller, T., Snyder, K.: Energy storage devices for future hybrid electric vehicles. *Journal of Power Sources* 168, 2–11 (2007)
5. Dobner, D.J., Woods, E.J.: An Electric Vehicle Dynamic Simulation. GM Research Laboratories, pp. 103–115 (1982)
6. Salameh, Z.M., Margaret, A.C., Lynch, W.A.: A Mathematical Model for Lead-Acid Batteries. *IEEE Transactions on Energy Conversions* 7(1), 93–97 (1992)
7. Appelbaum, J., Weiss, R.: Estimation of Battery Charge in Photovoltaic Systems. In: 16th IEEE Photovoltaic Specialists Conference, pp. 513–518 (1982)
8. Baudry, P.: Electro-thermal modelling of polymer lithium batteries for starting period and pulse power. *Journal of Power Sources* 54, 393–396 (1995)
9. Guerrero, M.A., et al.: Overview of Medium Scale Energy Storage Systems. *Compatibility and Power Electronics* 6 (2009)
10. Fuhs, A.E.: Hybrid vehicles and the Future of Personal Transportation. CRC Press, Boca Raton (2009)
11. Carp, M.C., et al.: Monitoring system and intelligent control system used in the starting process of a LDH1250HP locomotive. In: 12th OPTIM 2010, pp. 551–556 (2010)
12. Mitsubishi Electric Shows Prototypes of Ultracapacitor-Battery Hybrid Energy Storage Device (2010), <http://www.greencarcongress.com>
13. Ayad, M.Y., Rael, S., Davat, B.: Hybrid power source using supercapacitors and batteries. In: Proc. IEEE-PESC 2003, Acapulco (2003)
14. Cericola, D., et al.: Simulation of a supercapacitor/Li-ion battery hybrid for pulsed applications. *Journal of Power Sources* 195, 2731–2736 (2010)
15. Kötz, R., Carlen, M.: *Electrochimica Acta* 45, 2483–2498 (2000)
16. Conway, B.E.: Electrochemical supercapacitors – scientific fundamentals and technological applications. Kluwer Academic/Plenum Publishers (1999)
17. Borza, P.: Electric Power Cell. Patent EP 1796199 (2007)
18. Brice, L., Magali, L.: Automatic vehicle start/stop control method. US Patent 2009/0216430 A1, Dayton, OH US (2009)
19. Puscas, A.M., et al.: Thermal and Voltage Testing and Characterization of Supercapacitors and Batteries. In: OPTIM 12th 2010, Brasov, pp. 125–132 (2010)

Energy Management System and Controlling Methods for a LDH1250HP Diesel Locomotive Based on Supercapacitors

Marius Cătălin Carp, Ana Maria Pușcaș, and Paul Nicolae Borza

Transilvania University of Brașov, Electrical Engineering and Computer Science
29 Eroilor Street, 500036, Brașov, Romania

{ana_maria.puscas,marius.carp}@yahoo.com, paul.borza@unitbv.ro

Abstract. The present research is focused on developing solutions for increasing the lifetime of the batteries used on diesel locomotive. In order to increase the lifetime of the batteries and to reduce the fuel consumption and the pollutants, an intelligent starting system of the diesel locomotive is proposed. The starting system is composed of supercapacitors, batteries and adequate control system used for controlling the power flow from the storage devices to the internal combustion engine in the starting process. The implementation is described, modeled and simulated and the results of the simulations are compared with the experimental ones.

Keywords: supercapacitors, embedded control, energy management, diesel locomotive, lifetime.

1 Introduction

Lately, the population became more and more dependent on transportation. Even if we speak about public transportation (buses, metros, railways) or personal vehicles, the trend of the fuel and energy consumption is alarming increasing.

Developing new solutions, technologies and methods for improving the energy efficiency represents a priority line in order to satisfy the actual energy requirements.

In transportation, to improve the performances of the storage devices researches are made. The new researches from nanotechnology field allowed developing new electric storage devices with increased performances, such as stacked supercapacitors (SSC) [1], [2], [3]. The transportation field had as target implementing different systems based on SSC that allow the displacement of the vehicles using the self energetic resources. Among such systems, it can be mentioned: (i) the subway from Moscow where an emergency system based on the energy stored on SSC and batteries allow the displacement of the metro train in emergency cases to the first station [4], (ii) the hybrid bus developed by NASA [5], (iii) diesel locomotives developed by institutes such as ELTechnology and "Werkstoffe & Technologien, Transfer & Consulting" [6], [7]. Even if there are multiple implementations related to this topic, this research still represents a thematic of scientific interest, the focus being oriented on increasing the lifetime of the batteries used in the starting process. Also, the research is focused on

offering major improvements of the power trains of the self drives structures thus increasing their energy efficiency. In present, the starting systems of the ICE based on a combination of batteries and SSC represents a major applications class.

The aim of this research is to improve the energy efficiency, the reliability and the performances of the actual starting systems implemented on the self drive mobile structures that use internal combustion engine (ICE). The goal of the research is to modify the architecture of the starting process of a classic LDH1250HP diesel locomotive (LDH) in order to increase the lifetime of the batteries, the energy efficiency and to reduce the fuel consumption and pollution. Thus, the present paper presents the design, the electric model, the prototype and the experimental results of the implemented system used for starting the LDH diesel locomotive.

2 Contribution to Sustainability

The present research significantly improves the actual starting systems of the ICE based on a combination of batteries of accumulators and SSC. The advantages of the prototype are related to reducing the number of the accumulators thus reducing the carried weight. By using combined solutions for storage devices (SSC - non polluting storage devices) the number of the starts and stops can be increased without having the performances of the batteries affected, thus reducing the carbon footprint. Thus, the running time of the LDH, the pollutant emissions and the fuel consumption can be significantly reduced. Also, by using the implemented starting system controlled by a microcontroller scheduler, the lifetime and the performances of the batteries and also the reliability of the ensemble are significantly improved. This prototype creates new opportunities in the field of power electronics used in the heavy transportation systems, increases the energy efficiency, increase the exploitation efficiency thus being in a good agreement with the principles of the sustainability.

3 Architecture of the System

In order to asses the goal of the present research, the architecture of the LDH was modified. The classic structure of the LDH is supplied from two banks of lead acid accumulators connected in parallel. One bank of accumulators consists in 8 12 V/160 Ah lead acid accumulators connected in series. In the new structure of the LDH, one of the two banks of accumulators (the second supplying branch) was replaced with three 12 F/110 V SSC connected in parallel, able to provide the equivalent power density of 218 kJ [8], [9], [10]. As consequence, the size of the replaced branch was reduced to half of its initial size and the power density was increased. Thus, the researched starting system is composed of lead acid batteries with high energy density and SSC with high power density for providing the high peak current pulses requested in the starting process, even when the ambient temperature is bellow zero centigrade thus increasing the lifetime of the battery.

4 Physical Implementation

The architecture of the prototype is illustrated in Fig. 1, the starting system in Fig. 2 and the control algorithm implemented on ECS and used for the starting system in

Fig. 3. The charging system includes current limitation and voltage regulator devices able to protect the batteries after the ICE is started. By switching on K3, the SSC is also charged at a constant current.

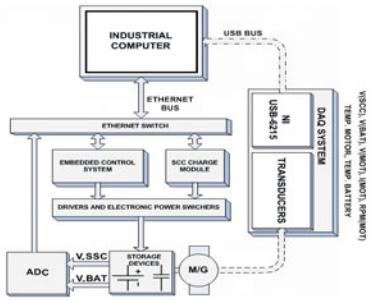


Fig. 1. General architecture of the prototype

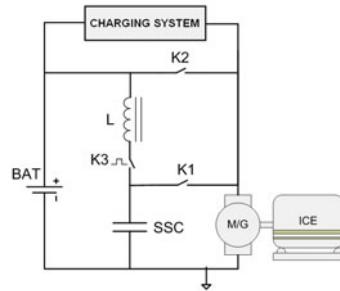


Fig. 2. Implemented starting system

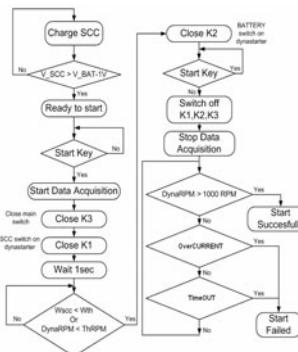


Fig. 3. Block diagram for starting the ICE of the LDH

The implemented and tested prototype consists in: (i) industrial computer, (ii) National Instruments data acquisition system (NI-DAQ), (iii) Ethernet network, (iv) electric high power devices (SCR) and (v) embedded control system (ECS) used for ensuring the correct power flow in the starting system.

As it can be seen in Fig. 3, before any start is performed, the voltage level of the SSC is automatically verified. If its voltage is below 96 V, the starting process will begin by charging the SSC from the battery at a limited current (K3 - switched on), thus protecting the battery. A flag indicates when the LDH can be started by the operator. After the starting button is pressed, the switches are firstly commuted on supplying the LDH from the SSC and after its energy is consumed, the starting process automatically commutes on battery in order to ensure the energy for maintaining the starting process. After the ICE is running, the switches are commuted thus ensuring the charging process of the SSC from the ICE. A voltage regulator was used in order to charge and to limit the voltage on the SSC and battery pack at 110 V, thus avoiding

the overcharging process which can damage the storage devices. If the velocity of the ICE does not reach its nominal value (1000 RPM) or the current absorbed from the battery is greater than a threshold current of 600 A, the fault situations can appear and an error is signalized. In the normal operation stage, the SSC play the role of a capacitive filter for smoothing the voltage on the electric circuits of the LDH. More than that, in case of switching off the locomotive, the SSC will maintain the maximum voltage level (around 96 V-110 V) for several hours. Thus, the voltage level on the SSC will facilitate the next starting process, the recharging of the SSC not being necessary anymore [2], [11], [12].

5 Simulations vs. Experiments and Results

Before implementing the system, to observe the behavior of the starting process of the LDH, simulations were made in PSpice tool. The model of the starting system is illustrated in Fig. 4 and the results of the simulations are illustrated in Fig. 5.

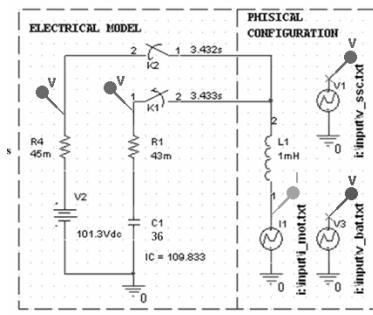


Fig. 4. Model of the starting system

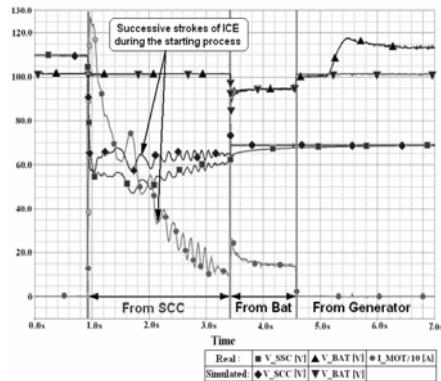


Fig. 5. Comparison between real and simulated experiments of the LDH starting system - battery and supercapacitor voltage and current

As it can be seen from Fig. 4, the model of the starting process of the LDH is composed of the 2 branches of storage devices (pack of batteries and pack of SSC) connected to the ICE through 2 switches (K1 and K2) electronically controlled. For modeling the packs of SSC and batteries, ideal models were used. The results of the simulations are compared with the real values of the current absorbed by the ICE in the starting phase.

As it can be seen in Fig. 5, the results of the simulations follow the experimental results. The current requested by the starting process of the LDH in the first 2.5 seconds is provided by the pack of SSC. A centrifugal mechanical regulator for controlling and maintaining the revolution speed at the desired value is used by the ICE. The centrifugal mechanical regulator automatically disconnects the ICE if the revolution is not maintained at its nominal rate for about 2 seconds after the starting was made. In order to ensure these requirements, the ICE is electronically switched to be supplied

from the batteries. As it can be seen from the experiments and simulations, the current provided by the battery in the starting process is about 5 times smaller than the current provided by the SSC ($I_{bat} = 16\% * I_{sc}$). Thus, the implemented system is protecting the battery from the high peak current pulses, they being provided by the SSC.

The difference in the behavior of the starting process between the simulations and results is due to the ideal models from PSpice used in the simulation process. To increase the accuracy of the simulation, complex models have to be used.

In Fig. 6 the calculated power and energy flow provided by the storage devices (SSC and batteries) to the load (ICE) can be observed. These values are determined by using the raw data –voltage and current – recorded with NI-DAQ. The ability of the SSC to provide peak power relative to the batteries can be observed. As result, in Fig. 6 the slope of the energy provided by the SSC (α) is grater than the slope of the energy provided by the pack of batteries (β).

In order to take the right conclusions, a data acquisition system mounted on the LDH, necessary to acquire the signals directly from the SSC, batteries and ICE was developed. Thus, an on - line software used for data acquiring and monitoring was implemented by using Visual Studio .NET tool.

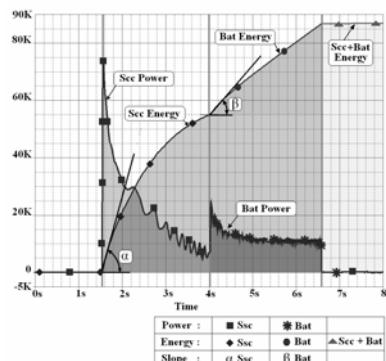


Fig. 6. Calculated power and energy corresponding to LDH starting process

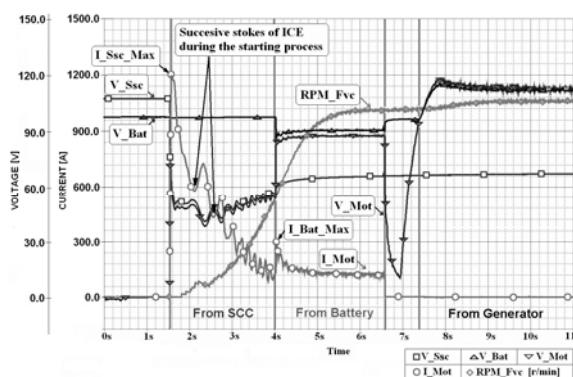


Fig. 7. Real records of the main parameters that characterize the LDH starting process

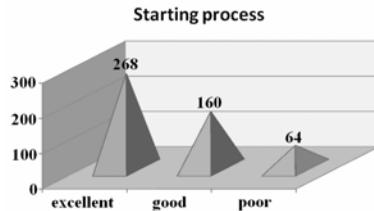


Fig. 8. Statistic of the starting processes of the LDH ICE

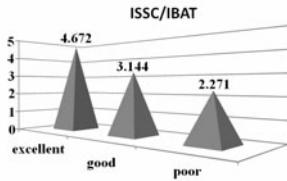


Fig. 9. ISSC/IBatt ratio

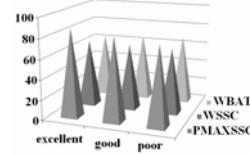


Fig. 10. PMaxSSc, WSsc, WBatt variation

The time variation of the voltages and currents - on SSC and batteries and also the revolution speed of ICE during a real starting process is illustrated in figure Fig. 7.

After defining several qualitative factors of the starting system, based on the acquired data, a statistic was realized. The chosen factors are: (i) the rate between the maximum current provided by SSC at the stating time and the maximum current supplied by the batteries at their commutation into the DC starting motor circuit (I_{SSc_Max}/I_{Bat_Max}), (ii) the voltage variation rate of the batteries at their commutation time on the DC starting motor, (iii) the maximum power provided by SSC and batteries during the starting process, (iv) the energy provided by the SSC and batteries during the starting process and (v) the revolution speed. All these data illustrate the accuracy of the starting process of the locomotive. We have also taken into account the temperature of the ICE and the ambient temperature. The real records reveal a normal variation of the above mentioned parameters. These data have been statistically analyzed and classified by labeling the starting processes as poor, good and excellent. The variation of the voltage on SSC at the initial moment was ignored because two situations were identified: (i) charging the SSC from the batteries as a result of a long resting period, (ii) charging the SSC from the generator of the LDH when the time period between the charging process of the SSC and the starting process of the LDH was short.

The statistic results are illustrated in Fig. 8, Fig. 9 and Fig. 10. From the total number of the starting process, only 1% was wrong and ignored. The successful rates of the starting process are illustrated in Fig. 8. From the total starting processes only 13 % were classified as being “poor”, and 87 % as being good or excellent. These statistic results validate our implemented and tested prototype.

The qualitative parameters mean values for the classified records are illustrated in Fig. 9 and Fig. 10. As it can be seen in Fig. 9, a good rate of the ISSC/IBAT ratio is

around 500 %. Thus, because of the significant reduction of the current provided by the batteries its reliability means a high protection of the batteries during the starting process of the LDH.

In Fig. 10 is illustrated the variation of the power and energy on SSC and batteries. The first series prove the dependency of the quality of the starting process to the power level injected by SSC during the first 2.5 s. The amount of the energy transmitted to the ICE is almost the same in all cases, thus emphasizing the importance of the power time dependencies offered by our system during the starting process.

6 Conclusions and Future Work

Usually, the performances and the lifetime of the batteries are affected especially by the high peak current pulses required by the ICE in the starting process. Because the implemented system is able to provide these peak current pulses from the SSC, the number of starts and stops of the LDH can be increased without the performances of the battery being affected. Thus, the size of the batteries was reduced to half of its initial value and the daily fuel consumption of the LDH used into the depot was reduced with around 6%. As consequence, the pollutants were also reduced. By using for the majority of the starting processes pre-charged SSC in previous displacements, the batteries are substantially protected, the biggest part of the power being supplied from the SSC.

The present paper describes the architecture of the prototype endowed with protection and self adapting features. Also, the algorithm detailed in the paper was used for controlling the correct functionality of the system and also contributed at increasing the performances of the starting system.

The paper also introduces a model of the starting process which was validated by the experimental results.

By using the experimental data acquired in 6 months (more than 1300 records), a statistic was made, thus determining the successful rate of the starting process. The results validated the prototype.

A set of criteria that reveal the accuracy of the starting process was defined and used for analyzing and classifying the experimental results. These criteria were applied for a collection of 492 samples from the total number of 1300 records. The simulations and experimental results clearly demonstrate the efficiency of the implemented system in the process of increasing the lifetime of the battery.

As future work, by identifying the real parameters of the system, a software sizing tool will be conceived. This toll is necessary to minimize the size of the batteries and supercapacitors by correlated these with the characteristics of the ICE.

Acknowledgments. This paper is supported by the Sectoral Operational Programme Human Resources Development (SOP HRD), financed from the European Social Fund and by the Romanian Government under the contract number POSDRU/6/1.5/S/6. Also, the present paper is a part of the national “TRANS-SUPERCAP” D21018/2007 project, currently under development at Transilvania University of Brasov.

References

1. Conway, B.E.: Electrochemical supercapacitors – scientific fundamentals and technological applications. Kluwer Academic/Plenum Publishers, New York (1999)
2. Chesa, A.: Locomotiva diesel hidraulica de 1250CP, pp. 382–390. ASAB Publisher (2001)
3. Sojref, D., Kuehne, R.: Supercapacitors – the European Experience in Transit Applications. In: Advanced Capacitor World Summit 2008, San Diego, CA (July 2008)
4. Sukhorukov, A.I.: Movement of metro train by supercapacitor - Electric train repair plant of Moscow metro. COST Action 542, Presentation
5. http://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19980013899_1998058116.pdf
6. <http://www.cantecsystems.com/ccrdocs/ELIT-Technology-Overview.pdf>
7. http://www.trecstep.com/German_Startup_Diesel_Start_engl.pdf
8. Petreus, D., et al.: Modeling and Sizing of Supercapacitors. Advances in Electrical and Computer Engineering 8(2), 15–22 (2008)
9. Diab, Y., et al.: Self Discharge Characterization and Modeling of Electrochemical Capacitor Used for Power Electronics applications. IEEE Transactions on Power Electronics 24(2), 510–517 (2009)
10. Bohlen, O., et al.: Ageing behavior of electrochemical double layer capacitors Part I. Experimental study and ageing model. Journal of Power Sources 172, 468–475 (2007)
11. Carp, M.C., et al.: Monitoring system and intelligent control system used in the starting process of a LDH1250HP. In: 12th OPTIM 2010, pp. 551–556 (2010)
12. Sojref, D., Borza, P.: Comparison of High-Voltage Supercapacitor Approaches and Case Study in Diesel Locomotive Starting System. In: 3rd European Symposium on Supercapacitors and Applications: ESSCAP 2008 (2008)

Home Electric Energy Monitoring System: Design and Prototyping

João Gil Josué, João Murta Pina, and Mário Ventim Neves

Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa

Monte de Caparica, 2829-516 Caparica, Portugal

jgj@netcabo.pt, jmmp@fct.unl.pt, ventim@uninova.pt

Abstract. The energy resource management is a major concern worldwide. Energy management activities minimize environmental impacts of the energy production. Therefore, electric energy consumption monitoring has been proposed as an important process which makes immediate reductions in energy use and CO₂ emissions. In recent years, advances in electronics have allowed the implementation of many technological solutions that could help to reduce energy consumption. This paper describes the design and prototyping of a home electric energy monitoring system that provides residential consumers with real time information about their electricity use. The system uses wireless communication and displays the information on a small LCD screen and on a computer.

Keywords: Electric energy monitoring, energy management, home automation.

1 Introduction

Electricity plays a crucial role in the economic and social development of the world and in the quality of life of its citizens and consumers [1]. However, electric energy production is mainly supplied by fossil fuels, such as oil, natural gas and coal. The dependency on limited fossil energy resources and the consequent greenhouse gases emissions (GHGs, including CO₂, CH₄ and N₂O) warned the world about the unsustainability of the current situation.

Investments in energy efficiency and renewable energy have been crucial measures of sustainable energy policies, since they are often economically beneficial, improve energy security and reduce local pollutant emissions [2].

In 2007, final electricity consumption in the residential sector in the European Union (EU) was 28% of the total [3]. This sector has been highlighted as an area which has a significant potential for improvement. Thus, residential sector energy efficiency programs can significantly reduce electricity consumption worldwide.

Consumers have an important role in the energy management activities and their actions represent an important step to minimize environmental impacts of energy production. Real-time electric energy information has a great impact in consumer's behaviors and habits. Past studies suggest that providing detailed and instantaneous feedback on household electrical demand can reduce electric energy consumption by 5-15% [4][5][6].

Due to advances in electronics and computing, many technologic solutions are now available. These solutions are a very important tool to a sustainable future.

This work describes the design, prototyping and testing of a home electric energy monitoring system capable of measuring and displaying on a small LCD the real time information about consumer's electricity use. This system can also be connected to a computer (via USB) to record and analyze measured data.

The developed system is part of a global energy monitoring system (electricity, gas and water) that provides real time energy use information to increase energy efficiency. Nevertheless, in this paper, only the electrical part is described, as this is able to work independently.

2 Contribution to Sustainability

Nowadays, the most common type of household electricity meter is the electromechanical induction meter. However, with this type of meter consumers have no means to judge electricity use other than their monthly utility bill. Therefore, it is difficult to know the necessary measures to improve the home's energy efficiency. The proposed system aims to contribute to sustainability in the way that it readily provides insight as to how and where the electric energy is being used. Since it provides real-time information on household electrical demand and records historical measured data for future analysis, this system potentiates electric energy efficient use, leading to significant environmental, political and economic benefits.

3 Home Electric Energy Monitoring Systems

Today, there are many systems available to monitoring household electricity use. Smart meters can be used to replace traditional electromechanical meters and provide both the supplier and the consumer with a better control of the electricity use. The smart meter can also be a part of energy management infrastructures, such as smart grids.

However, in a consumer's point of view it is beneficial to use a low-cost, user friendly and flexible monitoring system. This type of system, which is exclusively developed to help consumers manage and reduce their electricity use, doesn't replace the traditional meter and has the advantage of being portable. Depending on their sophistication, these monitoring systems can also communicate with a computer or a cell phone, send data throughout the house using wireless communications or connect to the internet allowing remote monitoring.

In recent years, several home electric energy monitoring systems have emerged on the market. Some systems can be plugged into the wall outlet to measure appliance's consumptions. Furthermore, other approaches measure total household electricity use through appropriate sensors installed into a home's switchboard circuits.

4 Proposed Monitoring System

The developed system consists of two electronic devices: data acquisition device and data display device, presented in figure 1. The data acquisition device measures

power and energy consumed by loads and the data display device displays measured data on a small LCD screen and sends results to the computer.

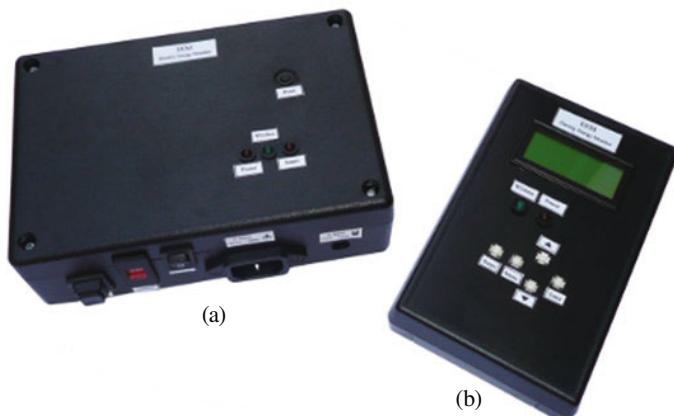


Fig. 1. Developed monitoring system's devices a) Data acquisition device b) Data display device

The main features of this monitoring system are: wireless communication between acquisition and display devices, monitoring capability at the appliance and switchboard circuit's level, average hourly energy use and electricity cost information display, and data recording on the computer.

Wireless communication between devices ensures greater flexibility and system's ease of use. The system's ability to monitor both appliance level and switchboard circuit's level informs the consumer about the balance of each appliance or circuit load. The knowledge of average hourly energy use and electricity cost provides important information that motivates changes to consumer's behavior. A computer connection is available to record measured data and to use it in many computer applications, such as, daily charting energy use data.

4.1 General Architecture

The data acquisition device diagram block is represented in figure 2. This device consists of five major blocks: power integrated circuit (IC), microcontroller, wireless transceiver, signal conditioning and power supply.

The data acquisition device measures line voltage and current signals through appropriate sensors. These analog signals are then conditioned and used by a power IC, which measure RMS voltage, RMS current, power factor and active power. This information is transmitted to a microcontroller that computes the energy consumed by a load and communicates with a wireless transceiver. The transceiver is responsible for sending the measured data to the data display device and for receiving commands from the user. The microcontroller can also communicate with status LEDs and with a reset button. The device power supply provides 5V DC from 230V AC line.

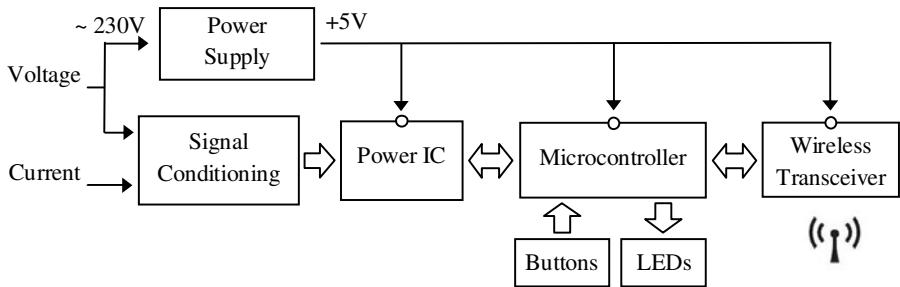
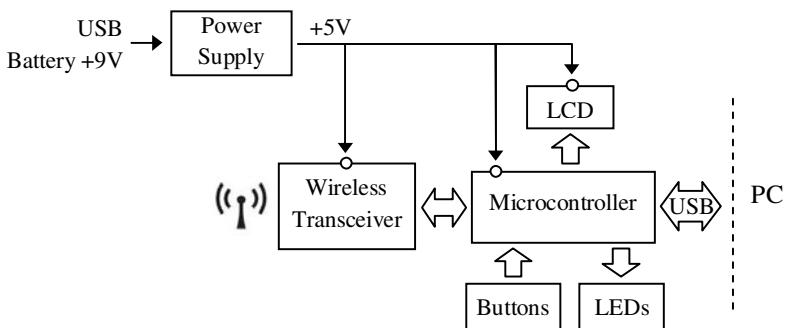
**Fig. 2.** Data acquisition device diagram block

Figure 3 presents data display device diagram block. This device consists of four major blocks: microcontroller, wireless transceiver, LCD and power supply.

**Fig. 3.** Data display device diagram block

Data display device receives measured data from the data acquisition device through a wireless transceiver. The transceiver sends the received information to a microcontroller. This microcontroller is responsible for several operations, including LCD data driving, communication with a computer via USB, LEDs and buttons control and electricity cost calculation. The device buttons allow the user to select and perform many functions, such as data measurement initialization, power IC calibration and electricity tariff definition. This device is supplied by a 9V DC battery or 5V DC USB interface.

4.2 Hardware

Power IC. The power measurement is performed by an analog integrated circuit design for residential single-phase, the CS5463 from Cirrus Logic. This IC focuses all the calculation complexity on a single circuit and is a highly accurate and a cost-effective solution. It is equipped with a computation engine that calculates RMS voltage, RMS current, active and reactive power and power factor within an accuracy of 0.1%. For communication with a microcontroller, the IC features a bi-directional serial interface which is SPI compatible.

Wireless Transceiver. Data communication between the acquisition and display devices is achieved by two wireless transceivers ER400TRS, developed by EasyRadio. These transceivers combine high performance low power RF, operate on 433 MHz frequency band, have a range of up to 250 meters line of sight, and provide serial interface for connection to UART devices.

LCD. The energy use information is displayed on a small LCD panel, developed by Batron, with 20 characters and 4 lines.

Microcontrollers. To control operations of many electronic components and endow the system with some intelligence, each device has an 8 bit microcontroller from Microchip. Data acquisition device is based on a PIC18F2420, which supports SPI and UART communication, while data display device uses a PIC18F2550 that provides a USB interface for communication with a computer.

Current Sensor. Current signal is read by a split core current transformer (CT) developed by CR Magnetics, model CR3110. This type of sensor enables the electricity monitoring at switchboard circuit's level, because there is no need to interrupt the circuit. CR3110 was designed to preserve linearity over a large current range, up to 65 A. Therefore, a burden resistor of $10\ \Omega$ was used for this system. This CT introduces a phase shift between the primary and secondary current. In this design, the phase shift was partially removed by the CS5463 throughout phase compensation.

Signal Conditioning. Voltage and current signals are conditioned by resistive networks and low pass filters to adjust their levels before being applied to the CS5463.

Power Supply. A transformer isolated power supply provides power to the data acquisition device. The AC from the center tapped secondary is full wave rectified, filtered and provided to the 5V regulator. The 5V loads are the CS5463, the PIC18F2420, the ER400TRS and the LEDs. Data acquisition device draws less than 47mA.

Data display device power supply comprises a 9V battery and a 5V voltage regulator to supply the PIC18F2550, the ER400TRS, the LCD and the LEDs. These components draw about 63mA.

Protections. To protect data acquisition device against possible surges, overloads and electromagnetic noise, a varistor, a circuit breaker and a line filter was used.

4.3 Firmware

The PIC18 microcontroller's firmware was developed in C and compiled with MPLAB C18 from Microchip. Each microcontroller (PIC18F2440 and PIC18F2550) need to perform many events. To provide real time response to these events, the microcontroller's interrupts and the interrupt services routines were used. A set of routines were developed for the two microcontrollers, which implement many real time events. The most important routines developed are listed below in table 1 and table 2.

Table 1. Data acquisition device PIC18F2420 routines

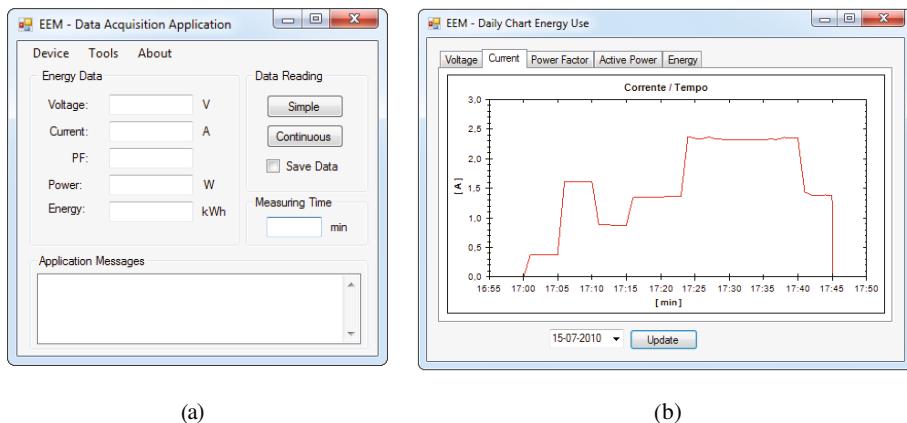
Routine	Function
UART receiver ISR	- Detect messages sent by the data display device
Timer 1 ISR	- Read CS5463 registers which have measured energy data; - Calculate energy consumed; - Send data to the data display device;
CS5463 calibration	- Eliminate CS5463 input channels offset

Table 2. Data display device PIC18F2550 routines

Routine	Function
UART receiver ISR	- Detect messages sent by the data acquisition device
Timer 0 ISR	- Update data displayed in the LCD
USB interface	- Control data transmission through USB interface
Port B	- Detect an input change in buttons corresponding pins

4.4 Software

A software application, based in C#, was also developed for the computer. This application communicates with the data display device through a USB interface and enables receiving and storing the monitoring system measured data. This program also provides daily charting energy use data giving a very useful visual tool to the user. Figure 4 presents some screens from the developed software application.

**Fig. 4.** a) Software application interface b) Daily chart energy use window

4.5 Prototyping

The developed system prototyping consists of three main parts: PCBs (printed circuit board) design and production, electronic components PCB welding and components boxes assembly. The prototype production cost was €280.82. To enable the appliance level monitoring, a module to be plugged into the wall outlet was also built. This module reads the line voltage signal and provides a wire for plugging the current sensor.

5 Results

In order to ensure a suitable data accuracy measurement, the proposed system was calibrated through a three-step process: CS5463 input channels offset elimination, phase shift compensation and scaling factors adjustment. To validate the system and verify its accuracy, several tests were carried out. These tests showed that data acquisition device has a measurement error of less than 1% for loads greater than 0,5A.

Finally, several functional tests were performed, to validate the system's behavior in real situations. The system was tested on both appliance level and switchboard circuit's level. It was possible to validate all the system's devices and components, which have a proper operation and integration.

6 Conclusions and Further Work

This paper describes the design and prototyping of a home electric energy monitoring system, which has been successfully completed. The experimental results have demonstrated that the proposed system is accordance with the design's specifications. Besides its original goal of integrating a global energy monitoring system, the developed system can also be used in energy auditing and energy advising processes.

Envisaged future work consists in new current sensor adding in order to double the data acquisition capacity, PCB's and component's size reduction by using SMT technology and internal memory installation to record measured data directly in the system.

References

1. Commission of the European Communities: Europe's current and future energy position Demand – resources – investments. Second Strategic Energy Review. p. 38. Brussels (2007)
2. Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K.B., Tignor, M., Miller, H.L. (eds.): Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge (2007)
3. Joint Research Centre (JRC): EU energy efficiency measures contribute to stabilize electricity consumption – drop in domestic use. Brussels (2009)

4. Mountain, D.: The impact of Real-Time Feedback on Residential Electricity Consumption: The Hydro One Pilot. Mountain Economic Consulting and Associates Inc., Ontario (2006)
5. Darby, S.: The Effectiveness of Feedback on Energy Consumption. Environmental Change Institute of University of Oxford, UK (2006)
6. Parker, D.S., Hoak, D., Cummings, J.: Pilot Evaluation of Energy Savings from Residential Energy Demand Feedback Devices. Final Report by the Florida Solar Energy Center to the U.S. Department of Energy (2008)

Sustainable Housing Techno-Economic Feasibility Application

Ricardo Francisco, Pedro Pereira, and João Martins

CTS, Uninova

Departamento de Engenharia Electrotécnica

Faculdade de Ciências e Tecnologia, FCT, Universidade Nova de Lisboa

2829-516 Caparica, Portugal

{r.francisco, pmrp, jf.martins}@fct.unl.pt

Abstract. The high prices currently achieved in the acquisition of non-renewable energy for electricity production and the low levels of energy efficiency in the housing sector are the national situation, which leads the Portuguese government to encourage the acquisition, the installation and the use of technologies which exploit indigenous and renewable energy. This study presents an application that was developed in order to help the citizen in his decision to invest in renewable technologies in their homes. The application is able to elaborate an economic analysis based on the selected type of renewable technology, providing the user with the knowledge of benefits and the annual costs involved in the system that he selected. This tool aims at facilitating the interaction of any user with such technologies and it can be used as a helpful tool to support the decision of investment in such systems.

Keywords: Renewable energy, energy efficiency, photovoltaic, solar thermal, rainwater systems, technical and economic feasibility.

1 Introduction

Energy consumption in buildings or homes (residential sector) is closely linked to the climate of the region where these are “installed” and depends on it, representing a major share of total energy consumption in most countries [1].

The number of dwellings in the EU-25 is about 196 million of which 80% are concentrated in seven countries, Germany 18.6% Italy 13.8%, UK 13.2% France 12.7%, Spain 10.8%, Poland 6.5% and Netherlands 3.5%. The entire building stock in the EU accounts for more than 40% of final energy consumption, of which 63% relates to final energy consumption in the residential sector [2].

Nowadays, people spend about 80 to 90% of their time inside buildings. Incorrect methods of design and construction lead to low energetic efficiency buildings or dwellings [3]. This low, or weak efficiency results in a large number of social problems, including problems of health and well-being and it also causes an excess of energy consumption which is responsible for high emission of pollutant and harmful gases into the atmosphere, particularly carbon dioxide (CO_2), and entail high

economic burden to their occupants [4]. It is therefore important the use of renewable technologies of energy in housing in order to reduce such costs and emissions and make it more efficient and sustainable.

With the increasing interest and knowledge regarding these technologies, we come to the need of tools capable of sizing and providing at the same time to the engineer, the designer, the user, and others economic perspectives that can help them in the decision to invest in these technologies. Such tools need to incorporate mathematical models of system components, know the possible situations that may occur in systems and rely on the weather information of various locations providing a pleasant and intuitive computing environment to the user [5].

There are currently several applications on the market able to perform simulations and provide the user with energy and economic analysis. However in the presence of the mentioned applications we've arrived at the conclusion that the majority relate to solar systems and are targeted at designers and engineers. They can be found at [5] and [6]. Concerning the systems of exploitation of rainwater, some applications have been found although they are mostly based on Excel spreadsheets. These applications can be found at [7], [8] and at the site of Sud Solutions company among others online. The computer tool developed differs substantially from these cases, as it is specifically intended for single family houses, geared to all users regardless of technical and specific knowledge they have about each technology.

This paper proposes a tool capable of performing a study of technical and economic feasibility of a sustainable housing in Portugal, being a sustainable housing the one which incorporates solar photovoltaic and solar thermal technologies – known as active technologies for decentralized energy generation - and technology of subsequent use of rainwater for non-potable uses or purposes, to provide the user interested in acquiring such systems for incorporation in his own house or any other, an initial idea of housing benefits and charges involved in these technologies and the possibility of comparing this type of housing with the standard one, i.e. without technologies using renewable energies. The application will consist of several sections, which encompass the choice of technology and data required for the proper functioning of it, the sizing of the systems and their economic analysis accounting for the economic gains and the energy savings associated with the choices made by the user as well as the periods or times of return (payback) of the investment in this type of housing.

This practice was developed based on various mathematical models proposed by different authors, who describe the individual performance of each technology, and it was developed in MatLab environment.

2 Contribution for Sustainability

The buildings or dwellings have shown as an area where environmental issues have not been taken in account at all. The reduction of energy consumption, of greenhouse gases, like CO₂, and of potable water consumption seems to be one of the key areas for achieving sustainability.

The aim is therefore to head for a higher level of energy sustainability, and to make this possible we must constantly take into account, that the way of achieving a more

sustainable "community" is through a reduction of energy consumption in buildings, and consequent reducing of pollution levels associated with the use of primary energy and its global influence on the climate, and the reducing of potable water consumption in buildings, by adopting active strategies that lead to higher levels of efficiency thus making housing more sustainable. In this sense, the developed application which is set out below has the following objectives: (i) to provide any user, regardless of their knowledge in such technologies, with the possibility of sizing systems using renewable energies by means of an estimate of energy and of rain water usage over a year, (ii) to provide regardless the user's knowledge in such technologies, an initial idea of costs and benefits he could reach with the installation of such systems in his home and (iii) to offer a full economic analysis to the selected system by comparing it with a housing that does not have any of these technologies.

This application aims at supporting the decision to invest in systems using renewable energies focusing on their interconnection with the final user. Figure 1 illustrates the framework of the conceptual model of the application.

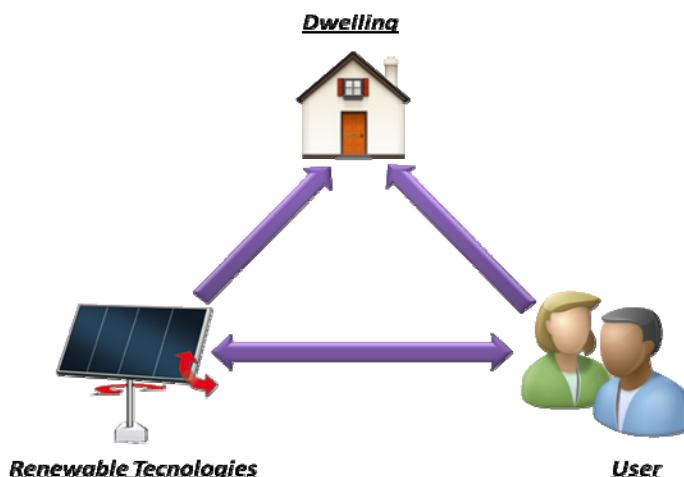


Fig. 1. Framework of the conceptual model of the application

This approach represents the base of the structure and the way the application was conceived, always taking into account the user's needs and requirements. Through a graphical interface, the user can visualize and control the necessary variables that will bring him to an outcome that may or may not meet his expectations.

In general terms the "macro" structure of the application, illustrated in figure 2, is based on two layers. The first regards the layout of the application and the interaction with the user and the second regards selection / command and control. From the second layer the user can access to other secondary layers that correspond to each of the technologies that make up the application and subsequently to its sizing and its economic performance.

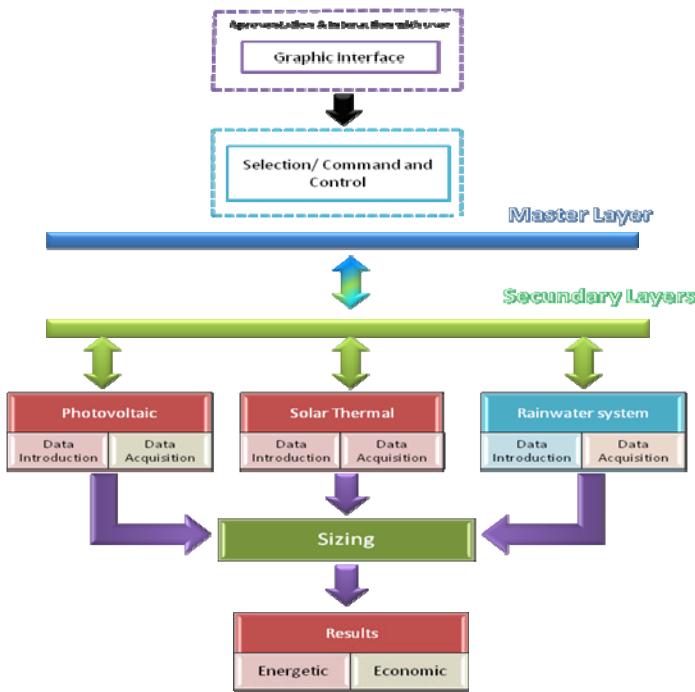


Fig. 2. Adopted structure in the application development

3 Application Structure

The adopted structure in the development of this application aims at facilitating the interaction with the user. To accomplish this goal, it has been divided in several blocks of which the selection / command and control, the introduction the acquisition of data, illustrated in Figure 3, and the graphical interface are to be distinguished and explained below.

Selection/command and control

The selection / command and control is responsible for the orders given to the application via the graphical interface to call the layers for each technology as well as the layer for the energetic calculations. The selection / command and control can be viewed as the desktop of the application once all the others are invoked from here, with the exception of the layer related to the economic calculations that is released from the sub-layer of energy calculations and from the layers related to climatic data, released from the secondary layers of each technology.

Introduction / data acquisition

The data entry can be done in two possible ways, either by user or by reading and importing data from a database (data acquisition). There are common data which have to be imported such as solar radiation and environmental temperature for the photo-voltaic and solar thermal systems and meteorological data for the rain water system

and specific data for each system. All data were collected and inserted in a database consisting of an Excel file. This database was compiled as follows: The temperature and radiation data were acquired online at the website of the European Commission through the free application Photovoltaic Geographical Information System (PVGIS) and meteorological data acquired through the Portuguese website of the Institute of water through the national information system of water resources (SNIRH). It also collected data regarding the technical components of each system (technical characteristics of pv modules, inverters, solar collectors, hot water storage tanks and storage tanks for rainwater) and their prices. These data were provided by companies currently operating in the Portuguese market. The graphical interface allows the user to select a particular component, depending on the selected system and the application performs a search of the database and imports the characteristic data for the selected component. The import of the climate data is performed in the same way taking into account the available information of the different cities and their municipalities in Portugal delivered to the user.

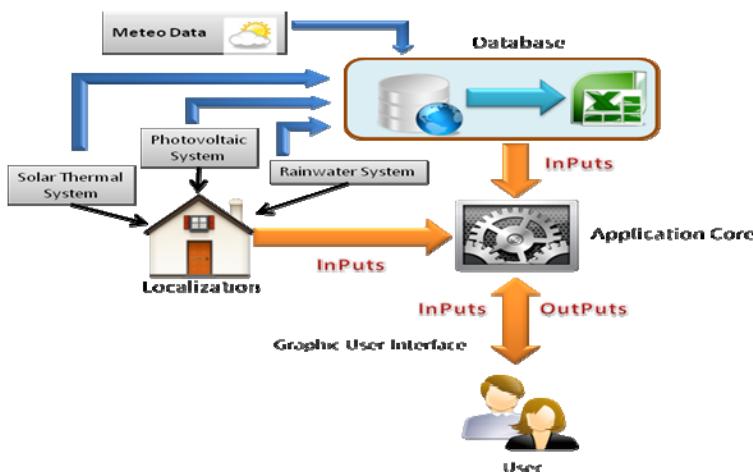


Fig. 3. Schematics of the Introduction / Command and Control

The core of the application is where you process all requests, orders and necessary responses, i.e., it acts as a kind of coordinator of services. It coordinates the delivery of services in order to provide an adequate response to a particular user's need, which may involve the selection of particular equipment, the associated calculations and the visualization of the results.

Graphic interface

The GUI shown in Figure 4 is responsible for the user interaction, helping him in the choice of several possible options for each technology and in the visualization of the data depending on his choice and the out coming results. It is equally responsible for the acquisition of necessary data which are demanded to the user. It was developed with the aim of providing the user with a clear and friendly environment.



Fig. 4. Graphical interface, the main application window

In Fig. 4, the main interface is divided into several areas, each one related to a specific task. A summary of each block, identified with capital letters, is presented below:

A - represents the acquisition module of the housing location, essential for the proper acquisition of climatic data for solar technologies and of rainfall data associated with the technology of rainwater harvesting.

B - represents the module of the available renewable technologies for selection.

C - represents the module for launch of energy calculations that can be seen as the sizing module of the system previously selected by the users for integration in his home.

By means of the main window, the user selects the city and the municipality where he is and then can decide later which technologies he wants to install. You can either select one, two or all three technologies. The choice of technology leads to a new window, where the necessary data for proper sizing of each system for subsequent economic analysis will be inserted and acquired. At the end of the introduction and acquisition of necessary data, there is a simple scheme of the connection of the individual components and of the overall system for better understanding by the user of the selected system. After the introduction and acquisition of all necessary data for each system, the user is introduced through the main application window to the next step, illustrated in Figure 5, which includes the step of sizing systems and the path to the final step, the economic evaluation of the entire system selected. This window is available to the user to configure the systems, which form a basis for economic analysis and the estimation of energy produced annually, and / or potable water saved annually. It is a window of "calculation", i.e. it does not require any insertion of data; the user has only access to the button for each technology, which calculates all necessary parameters and inserts them in the respective fields for later viewing. Next you can then follow the final step, illustrated in Figure 6, where after entering some data, such as: discount rates, desired lifetime of the project, if the project will be funded or

not and if it is for how long, etc., you get the total investment needed for each technology and the total investment in all the selected systems, the annual benefits discounted to the period of life chosen for the project of each technology and of the global system selected, the return period on invested capital for the global system and the economic viability of the overall indicator selected by Net Present Value (NPV).

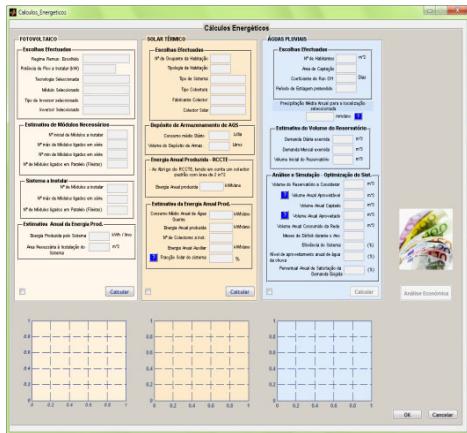


Fig. 5. System sizing

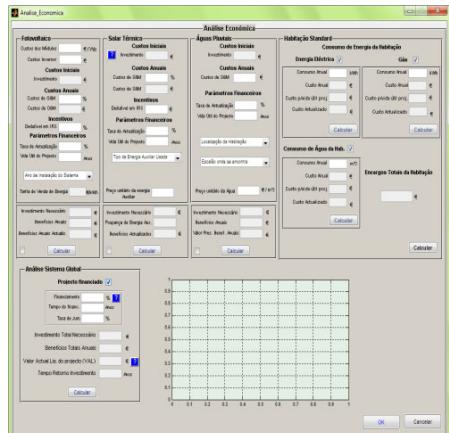


Fig. 6. Economic Analysis

4 Experimental Results

In this section, we present the results obtained for a residence located in the city of Aveiro and in the municipality of Vale de Cambra, which are illustrated in Figure 7 and 8. The economic analysis carried out corresponds to a system composed of the three available technologies in the application. A period of 15 years of useful lifetime for this project has been adopted, with a financing of 80% of the needed capital for 10 years, a discount rate of 7% and an interest rate of 6%. The choices for each technology are listed in Table 1.

The sizing of photovoltaic solar system resulted in the need for installation of 18 photovoltaic modules whose configuration results in nine modules connected in series along two rows. With this system you get an annual production of around 5500 kWh / year and you need an area of 23 m² for the system installation.

The solar thermal system sizing resulted in the need for a system composed of a solar collector and a storage tank of hot water with a capacity of 240 liters. With this system, the user can produce an energy saving of 1700 kWh / year corresponding to 53% of his energy consumption in water heating throughout the year and it requires an area of 2.50 m² for the system installation.

Taking into account the rainfall levels of the place, the sizing of the rainwater harvesting system resulted in a storage tank for rainwater with a capacity of approximately 20 m³. With the installation of this system the user has an available, usable volume of about 125 m³ of rainwater, of which he uses 112 m³ for the satisfaction of his dwelling consumption, i.e., he will consume annually about 81 m³ of potable

water from the network, leading to a system efficiency of around 58% having the system installed a level of utilization of 90% of rainwater which means that about 90% of rain that falls on catchment area of this place is saved by the system.

Table 1. Tecnology parameters

Photovoltaic		Solar Thermal		Rainwater
Remuneration scheme	Subsidized tariff	Housing typology	T3	Monthly demand (m ³)
Pp to install (kW)	3.45	Number of occupants	4	16
Solar module	Isofoton 180	Solar colector	AS-EFK 2.2	
Module type	c-Si			
Inverter	Sunny Boy 3800			
Inverter type	With transformer			

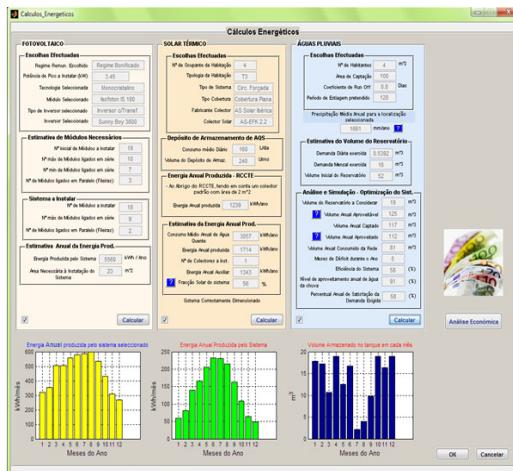


Fig. 7. Sizing of selected systems

As for the Economic Analysis, in the case of overall results, the total investment required is € 25,462. The update benefits during the lifetime of the project, amount to € 27,734, which means that the net present value (NPV) of the project is positive and amounts to € 2271. The decision to invest in this project can be considered favourable. Regarding the time of return on capital invested in this system, taking in account the financing and time allowed for financing, it will be paid within 11 years. Considering the adopted period of life 15 years, it means that the last four years of operation of the system will result in profit for you or for your home. This can be seen in Figure 8 through the graphic of the return period. Note that, in terms of cumulative value, the system will present, in the last year of life of this project, a benefit of approximately € 11,000.

Another interesting point to consider is the comparison between housing, with the integration of these technologies for exploiting renewable energy, and housing without them, considered as standard. In the analysis of housing considered as standard, an annual electricity consumption of 2600 kWh per year, a gas consumption of 650 kWh per year and water consumption of approximately 60 m³ per year are assumed. The same life period which was considered for housing with these technologies should be considered, in order to obtain a point of comparison. For the 15 years considered, the housing will present about € 18,170 of expenses in present value for this period. The housing with the incorporation of these technologies, as outlined above, presents a benefit for the same period of time amounting to € 27,734.

Then, establishing the comparison between the two houses it is clear that the option for sustainable housing, and taking into account the defined period of 15 years, the user can save in the end about € 9500, while continuing with its housing built without such technologies you will only have associated charges. It is also important to mention that by introducing this type of technology in his home the user makes it more efficient and sustainable.

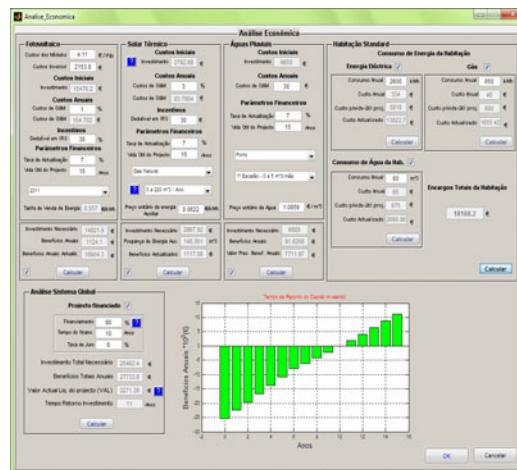


Fig. 8. Economic analysis of selected system

5 Conclusions

Introduction of renewable technologies of energy in housing makes homes more efficient and more sustainable and provides both economic and environmental benefits to its inhabitants also contributing to the security in energy supplying among others. Thus, this work has built an easy tool to use aiming at helping you in your decision to invest or not in these technologies.

A solid useful and accurate application resulted from this work providing results that can be displayed graphically (for a better perception) and be obtained from relatively simple data entry. It provides the user with a clear friendly graphical interface allowing him a most comprehensive knowledge of each renewable technology as well as the various aspects of its operation, either individually or at the system as a whole.

A consistent and profitable software tool in the supporting of investment decision in these technologies for incorporation into a dwelling is thus available for any user, regardless the knowledge he has about them.

Acknowledgments. This work was supported by FCT (CTS multiannual funding) through the PIDDAC Program funds.

References

1. Filippín, C., Larsen, S.F., Canori, M.: Energy consumption of bioclimatic buildings in Argentina during the period 2001–2008. *Renewable and Sustainable Energy Reviews* 14, 1216–1228 (2010)
2. Balaras, C.A., Gaglia, A.G., Georgopoulou, E., Mirasgedis, S., Sarafidis, Y., Lalas, D.P.: European residential buildings and empirical assessment of the Hellenic building stock, energy consumption, emissions and potential energy savings. *Building and Environment* 42, 1298–1314 (2007)
3. Pinheiro, M.D.: Ambiente e Construção Sustentável. In: Ambiente, I.d. (ed.) Fernandes \& Terceiro (2006)
4. Healy, J.D.: Housing Conditions, Energy Efficiency, Affordability and Satisfaction with Housing: A Pan-European Analysis. *Housing Studies* 18(3), 409–424 (2003)
5. Vera, L.H.: Programa Computacional para Dimensionamento e Simulação de Sistemas Fotovoltaicos Autónomos. Universidade Federal do Rio Grande do Sul (2004)
6. Argul, F.J., Castro, M., Delgado, A., Colmenar, A., Peire, J.: Edificios Fotovoltaicos: Técnicas y Programas de Simulación. In: Artes, S.L. (ed.) PROGENSA (Promotora General de Estudios, S.A.) (2004)
7. Bertolo, E.: Aproveitamento da Água da Chuva em Edificações. Faculdade de Engenharia da Universidade do Porto (2006)
8. Almeida, F.T.: Aproveitamento de água pluvial em usos urbanos em Portugal Continental: Simulador para avaliação da viabilidade. Instituto Superior Técnico, Universidade Técnica de Lisboa (2008)
9. <http://www.sudsolution.com>

Study of Spread of Harmonics in an Electric Grid

Sergio Ruiz Arranz, Enrique Romero-Cadaval, Eva González Romera,
and María Isabel Milanés Montero

Power Electrical and Electronic Systems, University of Extremadura, Avda. De Elvas,
s/n, 06011 Badajoz, Spain
sruiz@coeiex.es, {eromero, egzlez, milanes}@unex.es

Abstract. This paper presents an algorithm for estimating the Total Harmonic Distortion (THD) in the electrical power grid nodes, which is based on a load flow analysis by frequency component. The aim of this algorithm is, on the one hand, to model the linear and non-linear loads which would be part of an electrical network and, on the other hand, to estimate the THD which would appear in the network nodes, in order to evaluate its effects and consequences, as well as to choose the best alternative for solving these problems. The proposed algorithm is able to show the grid's buses in which the THD achieve a maximum, and could help to choose the most appropriate one to put the electronic devices up, as active power filters, optimizing the system design, permitting the reduction of the system losses and an increase of the energy transmission effectiveness into the grid.

Keywords: Non-linear Loads, Total Harmonic Distortion, Frequency Components, Power Flow, Power Losses.

1 Introduction

The electric energy consumers have a wide variety of electrical and electronic equipment that pollute the power grid, generating currents and / or voltage harmonics. In consequence, the operation of other users' equipment would be affected, due to the requirements of high quality power supply for proper operation (critical loads).

Harmonics are distortions or deformations of sinusoidal waves of voltage and / or current in electrical systems, mainly due to the use of non-linear loads (computers, televisions, variable speed drives, rectifiers, arc furnaces, fluorescent lamps, starters electronics, etc..), the use of ferromagnetic materials in electrical machines, switching operations in substations and in general the operation of switching equipment .

The effects of harmonics in power grids have been studied in previous literatures [1], [2]. To begin with, the appearance and the circulation of currents and / or additional voltages on the electrical system causes problems such as the increasing of active power loss, overloads in capacitors, measurement errors, malfunction protection, insulation damage, deterioration of dielectrics and decrease in the life of equipment, among others.

There are several alternatives for solving the pollution on the power supply through electrical or electronic equipments. Among these alternatives are active power filters.

These are electronic devices that have been studied and applied in recent years due to the advantages over other alternative solutions. It is generally assumed that the active filters will always work under the same specifications. This is a good consideration whether in reality the values of the parameters do not vary very atypical, so the filter will never show a pattern different from design.

2 Contribution to Sustainability

This paper presents an Harmonic Load Flow Algorithm to evaluate the Total Harmonic Distortion (THD) that appears in a electrical grid due to electrical and electronic components, obtaining the mathematical model of loads (linear and nonlinear) under the influence of harmonics, and could be used for evaluating the best alternative to put up the equipments (for example, active power filters) for solving the electrical pollution looking for the option that achieve the same goals but minimizing the total losses produced in the transmission and distribution grids. Also the algorithm determines the losses produced in the grid, that is a key parameter when operating and managing the grid and is one of the most important objective to be minimized for build a sustainable distribution grid.

This algorithm will help in future studies to characterize and model the demeanor of active filters, determining the nodes where these devices will succeed in improving the performance of the electric grid. Besides, this algorithm will help to evaluate the effects of these elements on mitigating the harmonics produced by non-linear loads and to estimate the decrease of electric grid losses. Only by using an algorithm of this kind it will be possible to analyze different active power filter strategies for comparing them and select the best depending on the electric grid characteristics.

3 Frequency Component Load Flow

The elements of an electrical system can be represented by linear or nonlinear impedances. The first case corresponds to those elements in which there is a proportional relationship between voltage and current to the same frequency range; on the other hand, non-linear elements do not have this proportional relationship across the spectrum. Among the elements that can be represented by linear impedances, are the lines, transformers, electrical machines and certain charges. On the contrary, the components which are considered as non-linear elements are mainly electronic equipment, like rectifiers.

3.1 Linear Load Modeling in the Presence of Harmonics

Transmission Lines have different mathematical models depending on their length, voltage and frequency. According to their length they are classified as short, medium and long lines.

They are fully described in [4], including their mathematical models. Thereby, the model which has been taken into account in the proposed algorithm is the π nominal one represented in Fig. 1a, where, r , l , c y g are, respectively, the resistance, inductance,

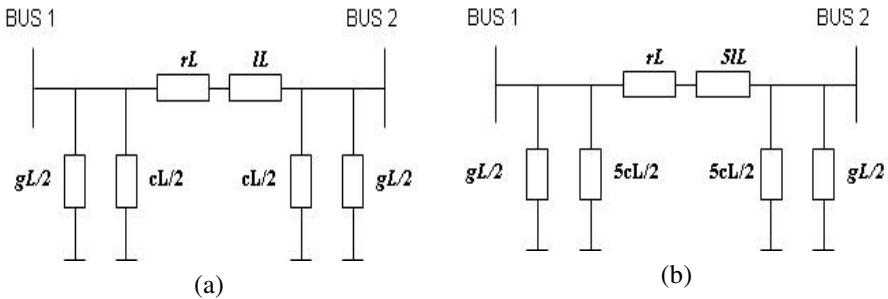


Fig. 1. (a) Equivalent single-phase model, (b) Fifth Harmonic line model

capacitance and conductance per length unit. According with [3], if the line is considered as a short line, c y g will be equal to 0. Figure 1b shows the equivalent single-phase model taking as an example the fifth harmonic.

Loads. In the study of harmonic flow, low-power loads are not represented individually; however, they are combined into equivalent circuits that represent the impedance characteristics of all charges. It is possible to consider variations in the impedance of the system due to the frequency or the chargeability level, both for domestic and industrial consumers. Nevertheless, as industrial loads are usually those that use capacitors for making up for the power factor, they are the ones that have more possibilities for contributing to the appearance of series and /or parallel resonance into the electrical system, [5]. They are fully described in [4], including their mathematical models.

Capacitors are modelled by their equivalent capacitance, providing a unique model that can be incorporated into a series or parallel circuit; these capacitive reactances must be multiplied by h [3], for taking into account the harmonic flow effects.

Non-controlled rectifiers (Fig. 2) are one of the sources of emission of harmonics in a power system. For modelling these loads it is necessary to take into account the distortion in the waveform in order to achieve a better description of the interaction with the network. According to the current injection model [3], the current wave is decomposed in Fourier series and each harmonic component is injected into the system as a power source; thus, it is possible to determine the system nodes harmonic voltages if a frequency sweep is made on the network. Equation (1) expresses the current injection model:

$$I_h = Y_{lh} \times V_1, \quad (1)$$

where Y_{lh} is interpreted as the relationship between each harmonic current and the fundamental component of tension, and is not necessarily linear.

Therefore, the nonlinear loads are modelled as constant current sources for each harmonic frequency and are calculated regarding to the fundamental frequency current.

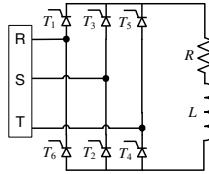


Fig. 2. Three-phase 6-pulse non-controlled rectifier diagram

These injections are based on the Fourier series, [4]. The Fourier series of AC input of a typical 6-pulse rectifier, when the output current is almost constant, fed by a transformer YY are listed below [3]:

$$I_{R,h} = I_{R,1} \times \left\{ \cos(\omega t) - \frac{1}{5} \cos(5\omega t) + \frac{1}{7} \cos(7\omega t) - \frac{1}{11} \cos(11\omega t) + \frac{1}{13} \cos(13\omega t) + \dots \right\} \quad (2)$$

where $I_{R,1}$ is the current drawn by the rectifier at the fundamental component, whose expression is obtained from:

$$I_{R,1} = \frac{4}{\pi} \int_{\pi/6}^{\pi/2} I_o \cdot \sin(\theta) \cdot d\theta. \quad (3)$$

By integrating the previous expression, the next one is obtained:

$$I_{R,1} = \frac{4}{\pi} I_0 [\cos \theta]_{\pi/2}^{\pi/6} = \frac{4}{\pi} I_0 \frac{\sqrt{3}}{2}, \quad (4)$$

and consequently:

$$I_{R,1} = \frac{2\sqrt{3}}{\pi} I_0, \quad (5)$$

Being

$$I_0 = \frac{V_0}{R} = \frac{1.35V_{L0}}{R} = \frac{1.35\sqrt{3}V_{FN}}{R}. \quad (6)$$

Finally, it can be shown from the expression (2):

$$|I_{R,h}| = \frac{I_{R,1}}{h}. \quad (7)$$

On the other hand, the equivalent resistance of rectifier for the purposes of modelling its behaviour for the fundamental component can be expressed by:

$$R_{eq,Y} = \frac{V_{FN,1}}{I_{R,1}}. \quad (8)$$

If the expressions (5) and (6) are used into expression (8), it is finally obtained:

$$R_{eq,Y} = R \times F_D \quad (9)$$

where:

$$R = \frac{(1.35 \times V_{LL})^2}{S_{RECTIF.}}; F_D = \frac{\pi}{1.35 \times 6} \quad (10)$$

3.2 Frequency Component Load Flow Algorithm

The frequency component load flow, like the conventional load flow, has different purposes. Firstly, to establish the state of the system taking into account the parameter of linear elements that shapes it. Secondly, to obtain information about the demanded power at the charge nodes and the generated power and, finally, to draw the topology of the system and the characteristics of nonlinear elements which cause the harmonics of voltages and currents that are multiples of the fundamental frequency in the system. The linear and nonlinear elements must be modelled considering the variation suffered with the frequency, according to previous epigraphs.

Once the conventional load flow is finished, consequently the models of linear and nonlinear loads have been defined. In order to find the harmonic voltages, the Y_{BUS} matrix should be built for each frequency and the next equation should be solved:

$$I_{(h)} = Y_{BUS}^{(h)} \cdot V^{(h)}. \quad (11)$$

In the above expression current injections are known due to their dependence on the nonlinear loads which are considered.

Once the mathematical model of the network elements are obtained, it is possible to make the flow of loads on the grid and, in consequence, to obtain the THD at each node, which is caused by the propagation of harmonics in the network due to the nonlinear loads, as well as to assess the power losses due to these harmonic currents.

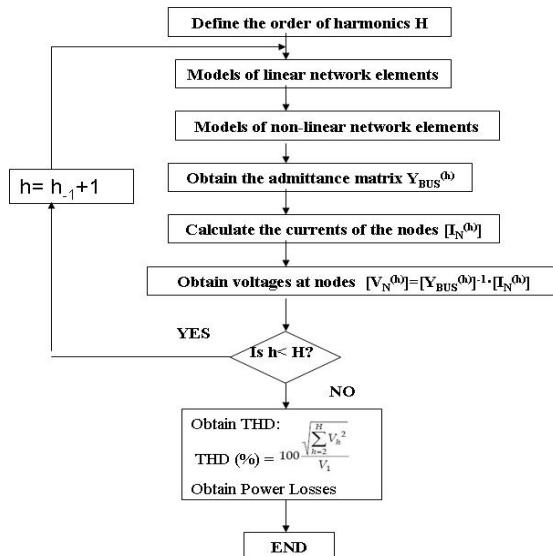


Fig. 3. Harmonic Load Flow Algorithm

Thus, the THD is the square root of the sum of the harmonic voltage amplitudes squared between the amplitude of the fundamental voltage component, which is expressed as a percentage. In mathematical terms, the THD is defined by the next expression:

$$THD(\%) = 100 \cdot \sqrt{\sum_{h=2}^H V_h^2} / V_1 \quad (12)$$

Where V_1 is the RMS fundamental voltage component, V_h is the RMS h -harmonic voltage component and H is the number of harmonic to evaluate.

Fig. 3 reflects the flow chart of the frequency component power flow algorithm.

4 Simulation Results

The proposed algorithm is applied to a radial power distribution line (Fig.4), which is based in the case presented in [7], in order to demonstrate its effectiveness for determining the THD in a power system. The circuit model of the radial distribution grid is illustrated in Fig. 4. The parameters in the simulation are included as follow:

- Power system: 220 V (line to line), 60 Hz. The transmission line parameters are $L_1 = 0.2 \text{ mH}$, $R_1 = 0.05 \Omega$, $C = 150 \mu\text{F}$, $L_2 = 0.4 \text{ mH}$ and $R_2 = 0.1 \Omega$.
- Nonlinear loads: two rectifiers rated at 2760 VA and 3328 VA are installed at bus 2 and bus 6, respectively. The dc side of the rectifiers consists of an inductor and a load resistor that model a load which produces harmonics.

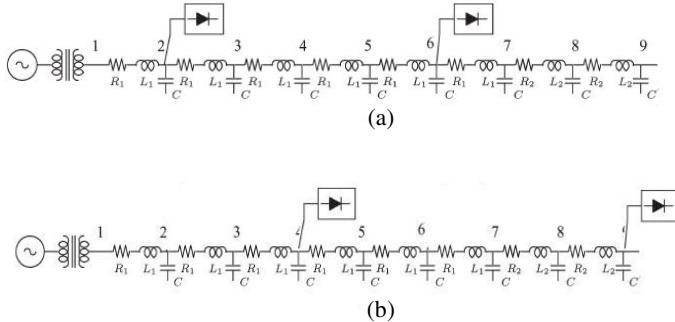


Fig. 4. Radial power distribution line: (a) Case I, (b) Case II

Network Parameters Expressed in the Per Unit System

The base power and the base voltage which have been taken into account for the per unit system are $U_b = 220 \text{ V}$ and $S_b = 6 \text{ kVA}$. As a consequence, the base impedance of the circuit is $Z_b = 8.0666 \Omega$.

The parameters of the radial network expressed in the per unit system are included inside the Table 1.

Table 1. Network parameters in the per unit system

Parameter	Real Value	p.u. Value
V_1	220 V	1
X_{L1}	$j\cdot 0.075398 \Omega$	$j\cdot 0.0093468$
R_1	0.05Ω	0.0061983
X_C	$-j\cdot 17.68388 \Omega$	$-j\cdot 2.192216$
X_{L2}	$j\cdot 0.150796 \Omega$	$j\cdot 0.0186937$
R_2	0.1Ω	0.01239669

Mathematical Models of the Lineal and Non-lineal Elements

Each of the elements of this system is modeled as follows:

- *Transmission line:* It is modeled using the Midline Model,[4] ,(see Fig.1). The conductance G is not considered in this network.

Rectifiers: They are considered as the non-linear elements of the network, which are located on Buses 2 and 6 (Case I), and buses 4 and 9 (Case II) respectively. According with epigraph 3.1, the parameters that define each rectifiers' mathematical model are the ones included in Table 2.

Table 2. Rectifiers in buses 2,4 and 6,9: Parameters in the per unit system

LOAD: RECTIFIER ON BUS 2 (I) and 4 (II)		LOAD: RECTIFIER ON BUS 6 (I) and 9 (II)	
PARAMETER	VALUE (p.u.)	PARAMETER	VALUE (p.u.)
S	0.46	S	0.554
R	3,9619	R	3,2897
F_D	0.27425	F_D	0.27425
R_{eqY}	1,5356	R_{eqY}	1,2759
$I_{Nn,1}$	0.3757	$I_{Nn,1}$	0.4525
$I_{Nn,h}$	$I_{Nn,1}/h$	$I_{Nn,h}$	$I_{Nn,1}/h$

Network Analysis and THD Calculation at Each Node

For this power system the simulations were performed using the MATLAB software. The results of the THD at each node for each case I and II are presented in Table 3.

Table 3. Total Harmonic Distortion in the network's buses

BUS	THD(%)	
	Case I	Case II
Bus 1	1.8038	1.4416
Bus 2	2.7262	2.1832
Bus 3	3.4746	1.8766
Bus 4	3.5541	1.9312
Bus 5	2.8873	2.8794
Bus 6	1.7368	2.8327
Bus 7	1.2930	2.9575
Bus 8	2.9950	2.5417
Bus 9	4.5654	2.3711

The Figure 5 shows the evolution of the THD on each node, which is based on the data collected in Table 3, and also show the losses of the transmission lines of the studied system, being 0.6209 p.u. for case I and 0.6048 p.u. for case II.

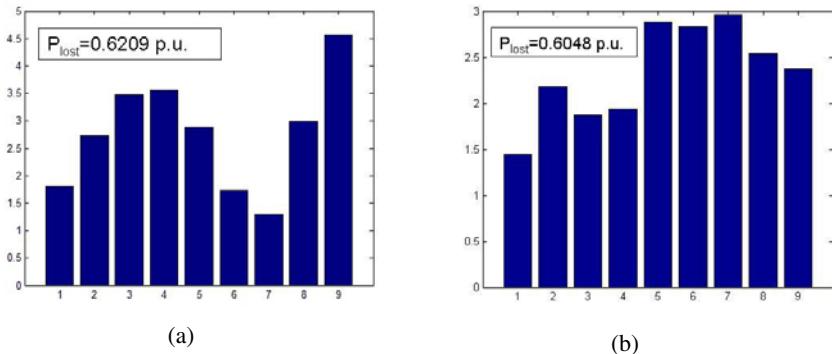


Fig. 5. Grid nodes THD for the system of Fig.4: a) case I, b) case II

5 Conclusion

In this paper an algorithm for estimating the Total Harmonic Distortion (THD) in an electrical power grid is presented. This algorithm is based on the harmonic load flow in a radial network, due to non-linear loads.

It has been shown by simulation that the proposed algorithm is able to show the grid's buses in which the THD is maximum, helping to choose the most appropriate one to put the electronic devices up, and optimizes the system design, permitting the reduction of the system losses and an increase of the energy effectively injected into the grid.

First of all, the different elements that could appear on an electrical network have been mathematically modeled, including transmission lines, linear and non-linear loads. Secondly, the Harmonic Load flow algorithm has been proposed, being able to calculate the Total Harmonic Distortion which appears on the network due to the non-linear loads. Moreover, the proposed algorithm has been applied to a radial power distribution line in order to demonstrate its effectiveness in predicting the THD in a power line. As a result, it could be established the system's buses in which the THD is maximum, helping to evaluate the best alternative to put up the equipments for solving the electrical pollution. Also this algorithm determines the losses in the distribution grid, and will allow, in future applications, to evaluate the influence of the presence of active power filter and the nodes where they are connected.

References

- [1] Arrillaga, J., Smith, B.C., Watson, N.R., Wood, A.R.: Power System Harmonic Analysis. John Wiley & Sons, Ltd., Chichester (2000)
- [2] Arrillaga, J., Watson, N.R.: Power System Harmonic, 2nd edn. John Wiley & Sons, Ltd., Chichester (2004)

- [3] Ríos Porras, C.A., Aristizabal Naranjo, M., Escobar, Z., Arrillaga, J.: Modelamiento de Sistemas Eléctricos en Presencia de Armónicos. *Scientia et Technica* 22 (August 2003)
- [4] Ríos Porras, C.A., Aristizabal Naranjo, M.: Modelamiento de Sistemas Eléctricos Y Empleo de Software Digsilent Power Factory en el Análisis de Armónicos, ch. 4. Technological University of Pereira (2001)
- [5] Ríos Porras, C.A., Aristizabal Naranjo, M., Gallego Rendón, R.A.: Análisis de Armónicos en Sistemas Eléctricos. *Scientia et Technica* 22, 21–26 (2003)
- [6] IEEE Working Group on Power Systems Harmonics: The effects of power system harmonics on power system equipment and loads. *IEEE Transactions on Power Apparatus and Systems* 104(9) (1985)
- [7] Cheng, P.-T., Lee, T.-L.: Distributed active filter systems (DAFs): A new approach to power system harmonics. *IEEE Trans. Ind. Appl.* 42(5), 1301–1309 (2006)

Impact of Grid Connected Photovoltaic System in the Power Quality of a Distribution Network

Pedro González, Enrique Romero-Cadaval, Eva González,
and Miguel A. Guerrero

Power Electrical and Electronic Systems (PE&ES),
School of Industrial Engineering (University of Extremadura)
<http://peandes.unex.es>

Abstract. Photovoltaic (PV) systems are increasingly present in the electrical distribution systems due to the governments incentives and low production costs of a developed PV technology. This paper summarizes the measurements on power quality (PQ) parameters carried out in a radial distribution network in two periods of time, before and after connecting a PV plant to the grid, and also shows the same parameters measured in the point of common coupling (PCC) of the grid and PV plant in order to discuss about how the impedance of the grid and ratio between injected power and power demanded by the load may influence changes on the PQ of the distribution system. Some measured values are compared with the limits set in the international standards. This paper assesses the impact of PV generation on the distribution system and important issues such as reverse power flow and harmonic distortion are analyzed.

Keywords: PV grid connected systems, power quality, distributed generation.

1 Introduction

The increasing number of photovoltaic systems in Spain is a fact in recent years due to the commitment made by the government with the European Union in terms of increasing the percentage of renewable energy in the generation mix. Optimal conditions for the development and implementation of this technology has meant that in 2008 6090 MW of PV power was installed in the world. Spain was the leader with approximately 43% according to [1]. Until recently, the performance of photovoltaic plants and technological improvements to reach an increase in the productivity, have been the focus of the research [2-5]. In short time, the advance of the solar technology is evident, and at this point it is necessary to deepen the performance of the facilities, but also to know the impact of those plants on the grid, even more because of actual PV high penetration levels in the distribution networks. The potential problems associated with high PV penetration levels are summarized in [6] and could be a disadvantage for the development of this renewable energy.

For unmanaged generation plants in Spain [7], the generation capacity shall not exceed one twentieth of the power network short circuit at the PCC in order to mitigate the possible effects of the PV plant on the distribution network. To obtain an accuracy value of the short circuit power in distribution networks before the installation of a PV plant, is not easy. The impedance of the grid, which is directly related to the short circuit power, changes with frequency and its estimation could be complicated with the presence of more than one generation source on the network.

PV system location on the distribution system could influence the PQ of the grid at the PCC [8] and also the difference between the load conditions and PV production could affects the voltage fluctuation of the grid due to the reverse power flow [9], [10]. Taking into account all these factors, this paper presents measurements performed on a distribution network in two situations, before and after the PV system connection to the grid with the aim of evaluate the impact of PV systems, not only in the substation, but also in the PCC.

2 Contribution to Sustainability

Grid interconnection of PV generation system has the advantage of more effective utilization of generated power with a more flexible and accommodated consumption. PV systems are a solution for the dependence and depletion of conventional energy sources and environmental problems. PV generation is increasingly widespread in the distribution network and quality problems have been detected that may affect the operation of the network. This paper presents experimental measures on a distribution grid, with and without connection of PV plant, to have a better understanding of the potential quality problems that this technology may introduce on the grid and to be able to solve them. Improving PV plants operation could increase the penetration level of PV plants in the distribution system.

3 Description of PV Grid Connected System

The size and the peak power of the PV system, the rated power and the short circuit power of the grid are important parameters to evaluate the PV influence on the grid, all of them related to the PV power penetration level. According to [11], most of the problems observed in an experimental analysis performed, occur in rural networks due to its high impedance. Taking this into account in the selection process of the line, measured data of different rural distribution lines was analyzed from year 2005 to year 2007, to choose one of them with a low load and not many changes in their annual load profiles. The line chosen has similar load profiles in different years with a low demand of power Fig. 1, and only reaches high values of load from April to August when rural customers develop their maximum activity. The other requirement to select the appropriate line was to have a recent PV generation connected to have measured data of both situations (with and without PV generation). The line studied is a rural distribution grid of 20 kV in the southwest of Spain with an approximated installed power of 5 MW that supplies to 115 customers, including a small town and also many dispersed irrigated farms. The line is connected to a 20 MVA transformer in a substation with two levels of voltages 66 kV/ 20 kV.

The PV plant is located 4 km from the substation and started operating in 2008. The PV system has 5 kW-2Φ, 20 kW-3Φ and 100 kW-3Φ inverters, all them equipped with AC galvanic isolation transformer and the usual protections of the majority of inverters used in PV installation such as: reverse polarity, AC over/under voltage, DC over voltage, AC and DC overcurrent, over temperature and antiislanding. PV park is made up of 6633 PV modules with 150 W of nominal power arranged in 390 parallel strings, with 17 modules in each. Strings are connected to inverters with different power. There are groups of 15 kW, 20 kW and 100 kW making a total of 995 kW of

installed power. All inverters are tied to the grid via four 0.4 kV/20 kV transformers, two of them of 630 kVA of power and another two of 400 kVA, which raise the voltage from 400 V to 20 kV before PV plant being connected to the grid.

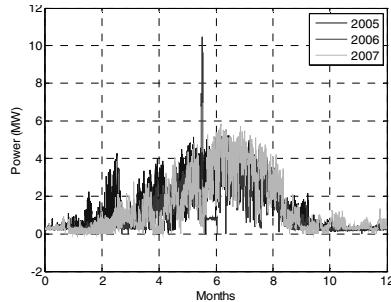


Fig. 1. Load profile of the grid from 2005 to 2007

4 Measure Methodology

Accomplishing with the objectives of this work, a PQ analyser (Topas 1000), was installed in two points of the grid: firstly at the substation, to obtain data measured in two periods of time (with and without PV generation connected to the network) and performance a comparative analysis of two monitoring periods and secondly at the PCC to evaluate the quality of the power injected into the network by the PV system.

Several electrical quantities and parameters as active and reactive power, power factor, total current harmonic distortion (THDI) and voltage harmonic distortion (THDU), individual current and voltage harmonics have been observed. All the monitoring periods were carried out for 24 hours in different days, grid without PV system was monitored at the substation on February 26, 2007; grid with PV system was monitored at the substation on March 31, 2009 and finally the output of PV system was monitored at the PCC on March 3, 2010.

PQ analyser Topas 1000 was connected at the low-voltage side of currents (300A/60A) and voltages (22kV/110V) transformers at the substation and in the same way, after the general switch of PV system to carry out the measurement at the PCC (Fig. 2). PQ analyser has 8 channels to measure currents and voltages; it is connected to four wire system with neutral-voltage/current and neutral common point connected to ground. Topas equipment recorded data with measurement interval of 10 ms for rms values, for this work, average intervals of 1 minute have been selected.

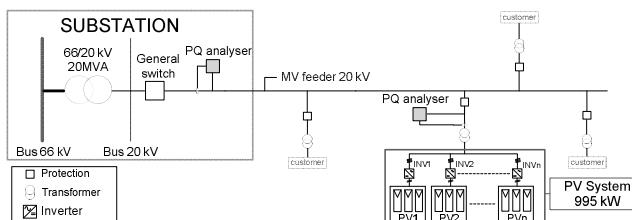


Fig. 2. Schematic block circuit diagram of the grid and PV system

5 Results

In the following subsections the most representative measured parameters are presented and compared with the limits set in the corresponding standards. [11] is considered to evaluate all values measured at the substation in 2007 (without PV plant) and 2009 (with PV grid connected plant). It refers to [12] for analyzing the values of the parameters measured at the PCC in 2010.

5.1 Power Quality on the Grid Measured at the Substation

It should be emphasized several issues to analyze the recorded data: the difference between the load and PV production significantly influences all parameters measured, that is; when the power produced by PV system is comparable with the power supplied by the main source of the grid, the behavior of the photovoltaic plant is most noticeable in the network, even more because PV system studied is near the substation and there is little impedance between them.

Active Power. The load profile of the grid in the last years was very flat and low, the value of 5 MW installed never is reached, even in the months of more activity. The demand of power monitored in both periods (2007 and 2009) was below 2 MW and on the other hand PV plant produced 995 kW in optimal conditions, so the reverse power flow could be. In Fig. 3 a), b); active power profile in 24 hours is shown, it is observed in b) how PV plant production affects the flow of energy for several hours at the substation.

Reactive Power. There was an increase in the reactive power consumption in the network with PV system Fig. 3 d) compared with the first period Fig. 3 c). The PV inverters are subject to the action of control systems aimed at providing zero reactive power at fundamental frequency, but several experiences has shown that the filters of inverters are not disconnected consuming reactive power, even when PV plant is not operating. This fact does not justify such a reactive power consumption, which may be due to increased loads on the grid in recent years.

Power Factor. As can be observed in Fig. 3 e), the power factor values are always above 0.85 and only fell below this number at night when the load profile is low, however in Fig. 3 f), power factor decreases to unacceptable levels during PV system operation. When PV system works with high power values close to rated ones, most active power demanded by the customers is supplied by the PV plant, reducing the demand of active power from the grid, but reactive power demand is the same, so it causes a low power factor measured at the substation.

Current and Voltage Harmonic Distortion. Voltage distortion is due to the currents demanded by nonlinear loads, these currents flowing through the impedances of the grid affecting the voltage nodes. The standard [11] limits THDU to 8% and also the individual voltage harmonics. Fig. 3 i), j) and Fig. 4 c), d) shows the voltage harmonic distortion does not exceed the limits set by the standard, so in terms of voltage distortion, the effect of PV system operation is negligible.

Current distortion is due to the current waveforms demanded by nonlinear loads and also the current injected by PV inverters into the grid. THDI Fig. 3 g), and also

high individual currents harmonics observed in Fig. 4 a) are due to a low value of fundamental component of current, however, it can be noted the presence of 40th harmonic due to switching signals of PV inverters at the substation Fig. 4 b), and extremely high values of THDI occur when PV system reach rated power and the reverse power flow is produced at the substation Fig. 3 h).

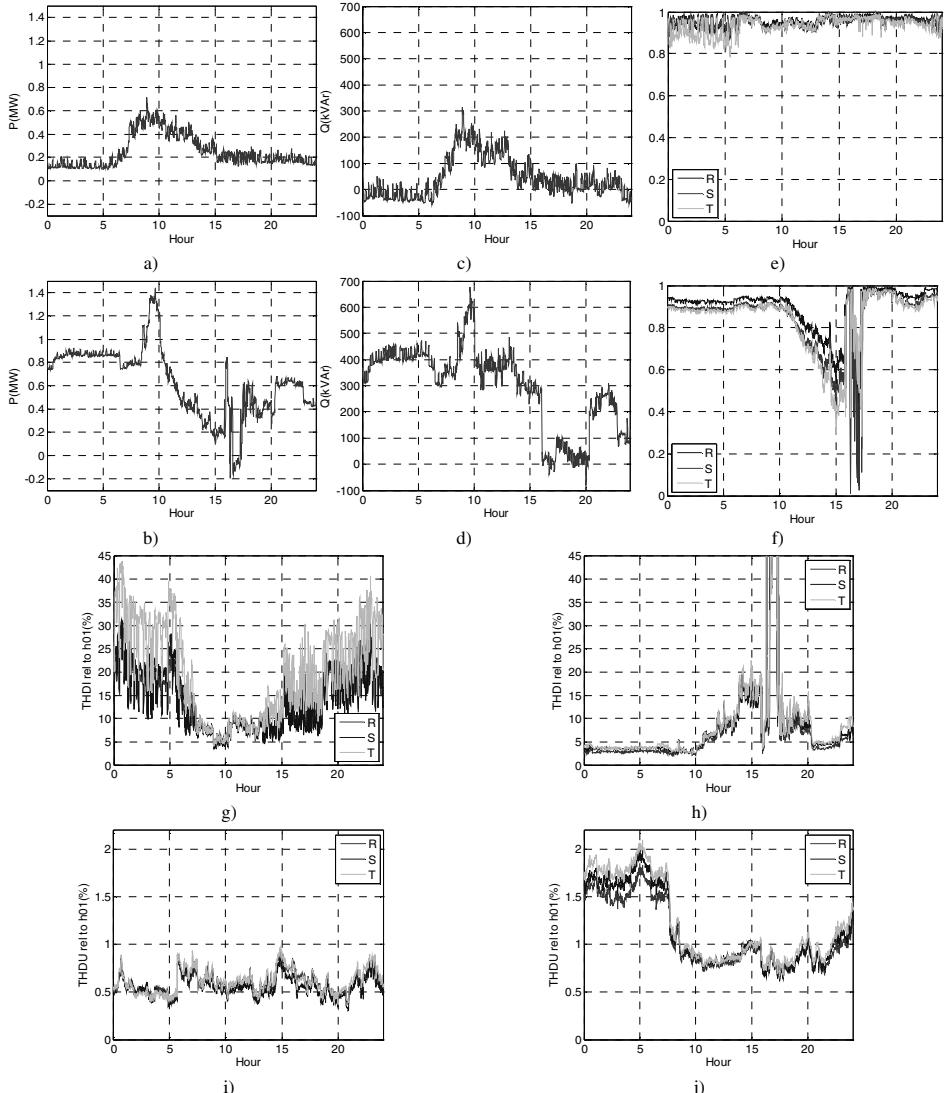


Fig. 3. Measures at the substation in 24 hours. Active power: a) without PV, b) with PV; reactive power: c) without PV, d) with PV; power factor: e) without PV, f) with PV; THDI: g) without PV, h) with PV; THDU: i) without PV, j) with PV.

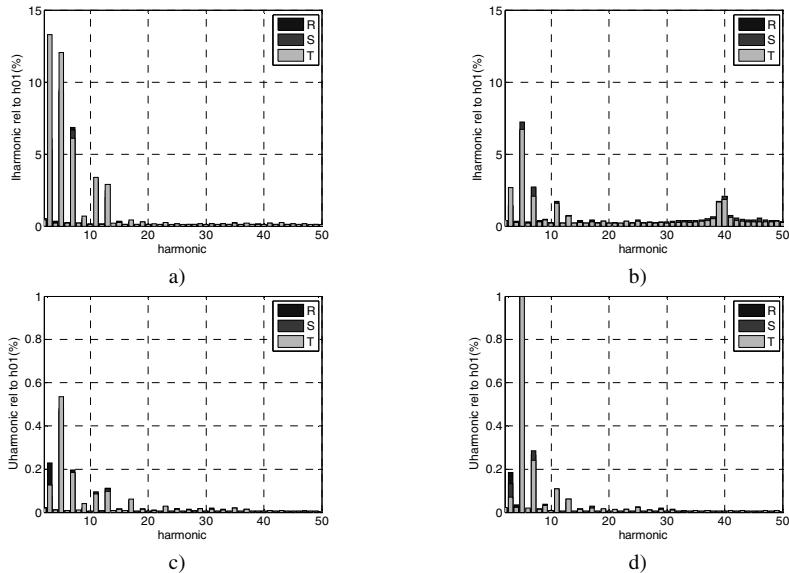


Fig. 4. Individual currents harmonics in 24 hours, a) without PV, b) with PV and individual voltages harmonics in 24 hours, c) without PV, c) with PV

5.2 Power Quality on the Grid Measured at the PCC

PV power quality injected into the grid is evaluated at this point, the rated power of PV system compared with the short circuit power at the PCC, is important to analyze the values of the different measured parameters and their influence on the grid.

In Fig. 5 a), PV plant active power profile is shown. The power fluctuations in active power profile are typical of a cloudy day and only the rated power is reached one time at 15 hour. Also in Fig. 5 b), is possible to see a demand of reactive power by the loads of the grid. It is observed fluctuations in reactive power before and after PV plant operation but there is not reactive power consume at night.

According to [12], the PV system should operate with a power factor above 0.85 when output exceeds 10% of the nominal power. As can be seen in Fig. 5 a), active power is always above 10% of the rated power of the plant (995 kW), however many times the power factor is below 0.8 Fig. 5 c). These values can only be justified if PV system provides reactive power compensation.

The voltage profile is shown in Fig. 5 d), it is observed that the values of voltage are within the normal voltage operating range set in [12]. Usually voltage variations are due to currents generated by the inverters that produce dangerous overvoltage when the PV power is similar to the power demanded by loads. The IEEE Standard 519-1996 states a maximum of 3% for the individual harmonic distortion and a maximum of 5% for THDU. None of those limits are reached at the PCC Fig. 5 f), h).

The current harmonic are related to the inverter operation, inverter manufacturers claim that their inverters provide a $\text{THDI} < 3\%$ when operating at 30 % of rated power and this situation is very common. The same occurs with the standard [12], it recommend the inverter to supply a current with less than 5 % TDHI when at the

nominal power. In Fig. 5 e), it can be seen that THDI measured is around this value. For individual current harmonic [12] limits in 4 % for 3th-9th harmonics, 2 % for 11th-15th, 1,5 % for 17th-21th and 0,6 % for 23th-33th. In Fig. 5 g), as can be observed all the limits are exceeded due to periods when the fundamental component of current is close to zero. It can be noted again at the PCC, the 40th harmonic due to switching signal of inverters, which is not set in the standards.

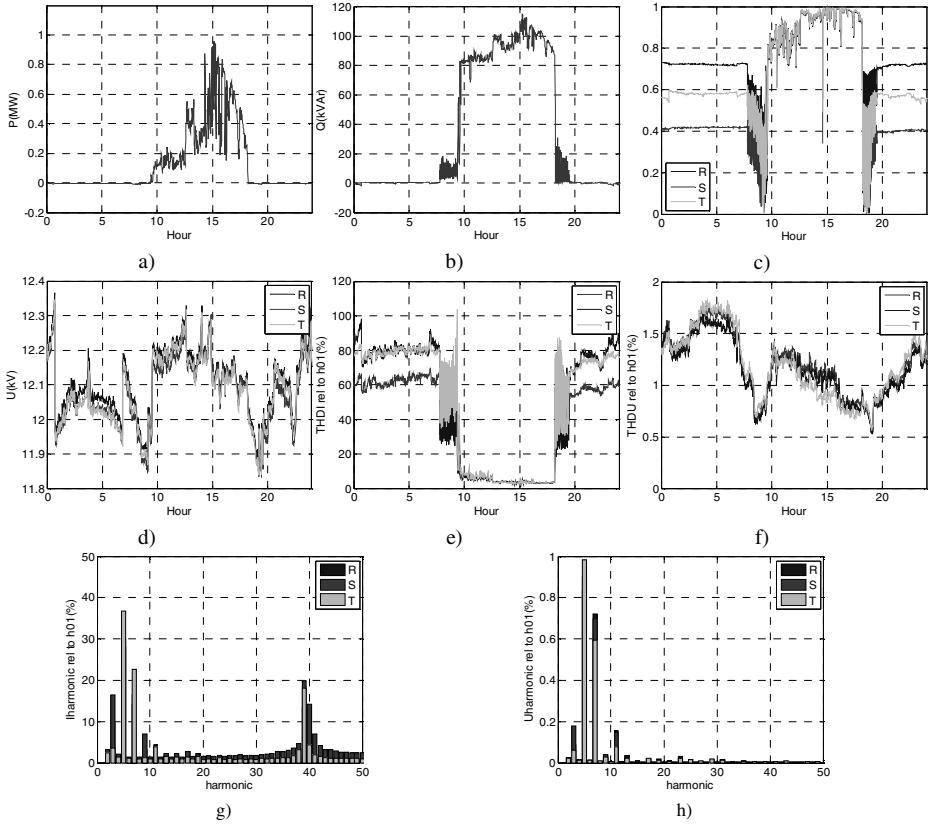


Fig. 5. Measures at the PCC in 24 hours: a) active power, b) reactive power, c) power factor, d) voltage variations, e) THDI, f) THDU, g) individual current harmonics, h) individual voltages harmonics

6 Conclusions

The measures carried out on the grid, have shown the influence of PV system in the power quality of the distribution network, the insertion of PV plant causes changes in measured quality parameters and operational particularities that do not seriously affect to the operation of the grid due to the power limit of twentieth of the power network short circuit for generation plants set in [7]. In a near future, if this power ratio increases, it could be necessary to find out solutions to mitigate the impact of these

quality particularities on the grid. Distribution network have been designed to operate in radial configuration with only consumption nodes, but actually it is common to find distribution networks with several generation points that could cause reverse power flow and voltage fluctuations in different parts of the grid. For this reason to review the management and protection devices of the distribution networks will be important issues. Storage devices and advanced inverters with internal controls that allow the adjustment of injected power as required by the grid, could be optimal options to taking into account.

Acknowledgements

This work was supported by electricity company Endesa under research contract with the University of Extremadura.

References

1. European Photovoltaic Industry Association publications: Global Market Outlook for Photovoltaics Until 2014 (May 2010), <http://www.epia.org>
2. Hun So, J., Seok Jung, Y., Jong Yu, G., Yeop Choi, J., Ho Choi, J.: Performance results and analysis of 3 kW grid-connected PV systems. *Renewable Energy* 32, 1858–1872 (2007)
3. Sidrach de Cardona, M., Mora López, L.l.: Performance analysis of a grid-connected photovoltaic system. *Energy* 24, 93–102 (1999)
4. Deb Mondol, J., Yohanis, Y., Smyth, M., Norton, B.: Long term performance analysis of a grid connected photovoltaic system in Northern Ireland. *Energy Conversion and Management* 47, 2925–2947 (2006)
5. Kymakis, E., Kalykakis, S., Papazoglou, T.M.: Performance analysis of a grid connected photovoltaic park on the island of Crete. *Energy Conversion and Management* 50, 433–438 (2009)
6. Eltawil, M.A., Zhao, Z.: Grid –connected photovoltaic power systems: Technical and potential problems-A review. *Renewable and Sustainable Energy Reviews* 14, 112–129 (2010)
7. Royal Decree 661/2007 on the Official State Gazette (BOE), 126, 22846-22886 (2007) Regulation of the activity of production of electrical energy in special regime
8. Srisaen, N., Sangswang, A.: Effects of PV grid-connected system location on a distribution system. In: IEEE Asia Pacific Conference on Circuits and Systems, Singapore, pp. 852–855 (2006)
9. Negrão Macêdo, W., Zilles, R.: Influence of the power contribution of a grid-connected photovoltaic system and its operational particularities. *Energy for Sustainable Development* 13, 202–211 (2009)
10. Canova, A., Giaccone, L., Spertino, F., Tartaglia, M.: Electrical impact of photovoltaic plant in distributed network. *IEEE Transactions on Industry Applications* 45(1), 341–347 (2009)
11. Voltage characteristics of electricity supplied by public distribution system, CEI EN 50160 (2000)
12. IEEE Std. 929-2000, IEEE Recommended practice for utility Interface of photovoltaic (PV) systems sponsored by IEEE Standards Coordinating Committee 21 on Photovoltaics, IEEE Std. 929-2000, IEEE, New York, NY (April 2000)

Power Quality Disturbances Recognition Based on Grammatical Inference

Tiago Fonseca¹ and João F. Martins²

¹ Faculdade de Ciências e Tecnologia, FCT, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal

² CTS, Uninova, Dep.^a de Eng.^a Electrotécnica, Faculdade de Ciências e Tecnologia, FCT, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal
`{tdf19282, jf.martins}@fct.unl.pt`

Abstract. A new approach in detecting and identifying power quality disturbances is presented. The use of Formal Language Theory, already exploited in other fields, was used to develop an innovative tool to identify patterns in electrical signals. The Concordia transform is applied to the 3-phase electrical system, composing a 2-D signal. The obtained signal is compared with a “healthy” 3-phase composed signal, retrieving new data. A Formal Language based inference algorithm is used to infer a grammar from this new data. Each type of fault has its own grammar, which allows the developed algorithm to easily detect and identify the disturbances.

Keywords: Power Quality, Formal Languages, Grammatical Inference, Pattern recognition.

1 Introduction

Monitoring power quality (PQ) has been an issue with great increase of relevance in the past years, part of which is due to the growth of electrical energy consumers, whether domestic or industrial. Most of today's loads are electronic based and therefore quite susceptible to disturbances. Occurrence of a power disturbance can cause several problems such as equipment malfunction/damage and losses in productivity. Power quality issues are addressed in IEEE 1159-1995 standard [1], where recommendations for PQ monitoring are also considered. Several monitoring methods analyse power line disturbances using different methodologies to detect and identify the disturbance. While some of them use mathematical morphologies [2] [3], there is also a strong emphasis on the use of multiresolution analysis such as the wavelet [4] [5].

In [6], voltage sag is defined as a brief decrease in the rms line-voltage as the voltage swell is an increase of the rms line-voltage. An interruption is a reduction of the line-voltage or current to less than 0,1pu. Voltage fluctuations are relatively small variations in the rms line-voltage. The variation in the 3-phase voltage amplitudes, relatively to one another, is described as a voltage imbalance. Harmonic distortion is the effect in the waveform by the existence of harmonics.

The research for a method that fits in all disturbances and computational requests is still in progress. Some methods work well on some disturbances and not so well on

others, where other methods have high processing costs. For example the wavelets approach needs to be decomposed many times to retrieve a significant conclusion.

This paper intends to be a contribution to this problem, presenting a new approach in detecting PQ disturbances. The objective is to detect and analyse 3-phase systems using grammatical inference learning algorithms. Formal language theory was initially presented by Chomsky [7] and has been used in distinct domains, such as detection of ECG signals [8], control of electrical devices [9], Chinese character recognition [10] or image parsing [11].

The basic idea is to infer one grammar for each type of PQ disturbances.

2 Contribution to Sustainability

The renewable de-centralized power production poses new and interesting problems concerning power quality. The quality of the delivered power often depends on the mechanical/electric/electronic power interfaces, their control strategies and its directly connected with the sustainability of the power delivery system. On the other hand, electronic loads are also a strong source of power quality disturbances. Due to their nonlinear nature, these loads inject harmonics current into the power system and cause voltage harmonics distortion. In order to keep the sustainability of the power delivery system it is important to understand how power quality disturbances influence the system's performance. An important issue is development of techniques capable of detecting and overcoming power quality disturbances in system-equipment interactions. The presented power quality detection method is a contribution for the power quality detection problem and thus for the sustainability of the power delivery system.

3 Formal Language Concepts

3.1 Grammar

The grammar of the language is a set of rules that specifies all the words in the language and their relationships. Once the grammar is found, the grammar itself is a model for the source. To define a grammar (G), one specifies a terminal alphabet, a nonterminal alphabet, a start symbol, and a set of productions.(1)

$$G = \{\Sigma_T, \Sigma_N, S, P\}. \quad (1)$$

- Σ_T , represents the terminal alphabet, a set of symbols that create words, where a word is a string of symbols.
- Σ_N Is the nonterminal alphabet, set of auxiliary symbols that will produce words by the production rules.
- S Being the start symbol, a special nonterminal symbol to start the production of words.
- P , productions are the set of substitution rules, creating words that fit in the specific language.

Example of a grammar representation:

$$\Sigma_N = \{S, A\} \quad \Sigma_T = \{a, b\} \quad P = \{S \rightarrow aS, S \rightarrow aA, A \rightarrow bA, A \rightarrow b\}$$

Retrieving the following language:

$$L(G) = \{a^n b^m \mid n \geq 1, m \geq 1\} \Rightarrow \{ab, aab, aaab, aabb, \dots\}$$

According to the Chomsky hierarchy, grammars are classified in 4 different types. (see Table 1)

Table 1. Chomsky's hierarchy of grammar types

Type	Name	Production Rules
0	Unrestricted	No restrictions
1	Context-sensitive	$\alpha A\beta \rightarrow \alpha\gamma\beta$
2	Context-free	$A \rightarrow \gamma$
3	Regular	$A \rightarrow aB$ $A \rightarrow a$

For this work only type 3 grammar are considered: regular grammars which can be represented by finite state automata.

The basic idea is that the use of the grammar allows the decision if a given word is part of the language defined by that grammar, being this the basis for the pattern recognition methodology presented in this paper.

3.2 Grammatical Inference

Grammatical inference is an algorithm that can identify the grammar from a set of positive and negative examples. Obviously the quality of the inferred grammar is directly connected with the quantity and quality of the learning examples.

In this work, the developed algorithm will search electrical signals samples for signs of the recursive rules that will characterise the grammar being sought. At the possibility of a recursion being identified, a step in the inference algorithm is completed by substituting the substring. The sample will be rewritten several times, each time the recursion will be substituted by a symbol based on regular expression theory. In the end the method returns an expression that represents the grammar inferred from the sample.

A simple example of the algorithm is presented below:

Considering the sample, $I^2 = (x_1, x_2, x_3)$

$$x_1 = aabaaababcabc \quad x_2 = abcabaabcbc \quad x_3 = aaaaabc$$

The sample is analysed to find recursive parts,

$$\begin{aligned}
 x_1 &= (a)^2 baaababcabc \\
 x_1 &= aab(a)^3 babcbc \\
 x_1 &= aabaa(ab)^2 cabc \\
 x_1 &= aabaab(abc)^2 \\
 x_2 &= abcab(a)^2 bcabc \\
 x_2 &= abcabaa(bc)^2 \\
 x_3 &= (a)^5 bc
 \end{aligned}
 \qquad \text{Possible matches: } a, ab, abc, bc$$

Choosing hypothesis: a . Rewriting: $z = a^+$

$$x_1 = zbzbczb \quad x_2 = zbczbcb \quad x_3 = zbc$$

Analyse:

$$\begin{aligned}
 x_1 &= (zb)^3 czbc \\
 x_1 &= zbz(zbc)^2 \\
 x_2 &= zbc(zb)^2 cbc \\
 x_2 &= zbczb(bc)^2
 \end{aligned}
 \qquad \text{Possible matches: } zb, bc, zbc$$

Choosing hypothesis: zb . Rewriting: $y = (zb)^+$

$$x_1 = ycyc \quad x_2 = ycycbc \quad x_3 = yc$$

One possibility: yc . Rewriting: $x = (yc)^+$

$$x_1 = x \quad x_2 = xbc \quad x_3 = x$$

Final expression, retrieved from the sample:

$$x + xbc = ((a^+b)^+c) + ((a^+b)^+c)^+bc$$

The choice of the alphabet is of extreme importance to retrieve useful results. As an example Table 2 contains the explanation of an alphabet that can be used in a ECG¹ signal piece (note that the fact that some symbols are in uppercase has nothing to do with non-terminal symbols and should be ignored in the scope of the example). Fig. 1 shows how the alphabet is applied.

Table 2. Primitives used for the alphabet, Δ is a minimum slope values specified beforehand, adapted from [12]

Primitive Name	Symbol	Description
Horizontal	h	A segment of constant value
Up slope	u	An upward segment with slope $< \Delta$
Down slope	d	A downward segment with slope $> -\Delta$
Strong up slope	U	An upward segment with slope $\geq \Delta$
Strong down slope	D	A downward segment with slope $\leq -\Delta$

¹ Electrocardiogram.

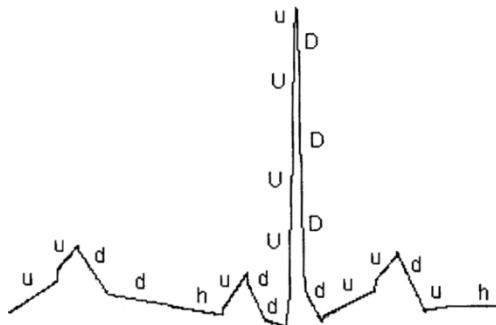


Fig. 1. Piecewise linear approximation of an ECG signal [12]

4 Implementation and Results

4.1 Preparing the Electrical Signals

Since we are working with 3-phase systems, Concordia transform was chosen to get the presentation of the electrical signal in 2-D space. By applying the Concordia transform to an undisturbed electrical signal and to a disturbed one, it's possible to retrieve a new representation, which consists of the difference between the two transformed signals. Fig. 2 presents a disturbed electrical signal (voltage fluctuation) and an undisturbed one.

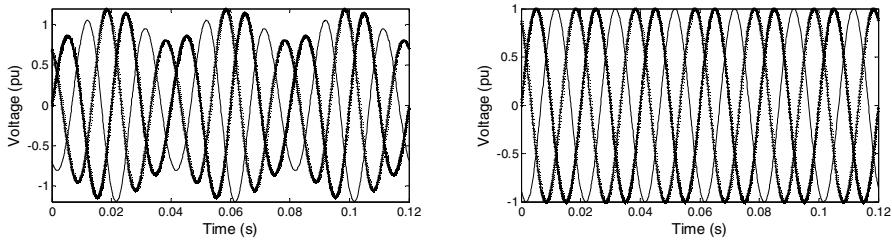


Fig. 2. Disturbed and undisturbed signal

Since the analysis is on a 3-phase system, applying the Concordia Transform makes it possible to analyse the system in a 2-D scenario. Implementing the transform into each signal and overlapping the results, one gets the signal shown in Fig. 3: a perfect circle corresponding to the normal signal and another one corresponding to the faulty signal product. Determining the radial distance between both results along the signal sampling, the outcome is presented in Fig. 4. Each type of PQ disturbance grammar will be inferred from this composed signal.

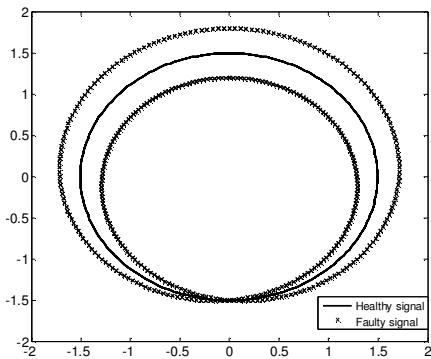


Fig. 3. Overlapping transforms

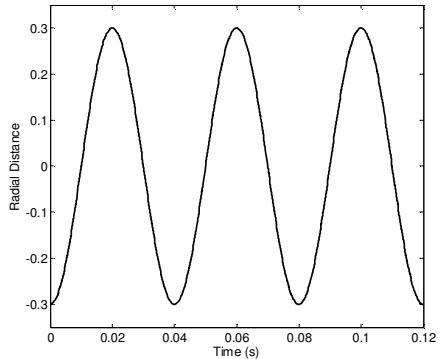


Fig. 4. Time evolution of the radial distance

4.2 Inferring the Grammar

As stated before, the chosen alphabet will highly affect the final result in the inferred grammar. For this work a 4 level alphabet, $\{a,b,c,d\}$ was chosen. Applying this alphabet to the composed signal presented in Fig. 4, one obtains the word sequence, as exemplified in Fig. 5 with a random signal.

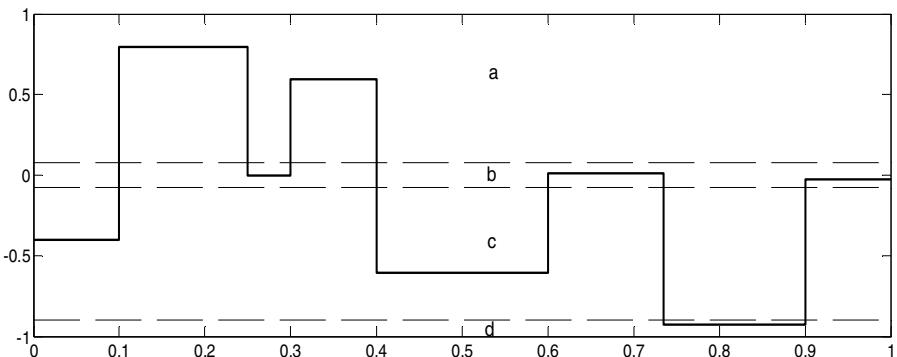


Fig. 5. Representation of the used alphabet

Using the proposed methodology in different disturbances of the same type, gives us a sample of strings to analyse and infer the respective grammar. The obtained grammars for each disturbance are presented below, with the pattern that characterises each disturbance (showed in the regular expression form):

- Voltage imbalance $((bc)^+(ba)^+)^+$
- Harmonic distortion $(abcb)^+$
- Voltage fluctuations $(cbab)^+$

- Interruption *bcd*
- Sag *bcb*
- Swell *bab*

5 Conclusion

This work presented the development of a novel algorithm based on the use of formal languages in detecting power quality disturbances. The grammatical inference algorithm was developed in order to retrieve the grammars that characterise each disturbance, being each disturbance characterised by one, and only one, grammar. Being the grammars established is possible to detect and classify the disturbances, analysing them by means of an automata or inferring the grammar of a given signal.

References

- [1] IEEE Recommended Practice for Monitoring Electric Power Quality. IEEE Recommended Practice for Monitoring Electric Power Quality (1995)
- [2] Radil, T., et al.: PQ Monitoring System for Real-Time Detection and Classification of Disturbances in a Single-Phase Power System. *IEEE Transactions On Instrumentation and Measurement* 57, 1725–1733 (2008)
- [3] Matz, V., et al.: Automated Power Quality Monitoring System for On-line Detection and Classification of Disturbances, pp. 1–6. IEEE, Los Alamitos (2007)
- [4] Gaing, Z.-L.: Wavelet-based neural network for power disturbance recognition and classification. *IEEE Transactions on Power Delivery* 19, 1560–1568 (2004)
- [5] He, H., Shen, X., Starzyk, J.A.: Power quality disturbances analysis based on EDMRA method. *Electrical Power and Energy Systems* 31, 258–268 (2009)
- [6] Kusko, A., Thompson, M.T.: Power Quality in Electrical Systems. MacGraw-Hill, New York (2007)
- [7] Chomsky, N.: Aspects of the theory of syntax. MIT Press, Cambridge (1965)
- [8] Trahanias, P., Skordalakis, E.: Syntactic pattern recognition of the ECG, vol. 12, pp. 648–657. IEEE Computer Society, Los Alamitos (1990)
- [9] Martins, J.F., et al.: Language Identification of Controlled Systems: Modeling, Control and Anomaly Detection. *IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews* 31, 234–242 (2001)
- [10] Kuroda, K., Harada, K., Hagiwara, M.: Large Scale On-Line Handwritten Chinese Character Recognition Using Improved Syntactic Pattern Recognition, vol. 5, pp. 4530–4535 (1997)
- [11] Han, F., Zhu, S.-C.: Bottom-up/top-down image parsing by attribute grammar, vol. 2, pp. 1778–1785 (2005)
- [12] Marques, J.P.: Pattern Recognition, Concepts, Methods and Applications. Springer, Heidelberg (2001)

Weather Monitoring System for Renewable Energy Power Production Correlation

Marcos Afonso, Pedro Pereira, and João Martins

CTS, Uninova

Departamento de Engenharia Electrotécnica

Faculdade de Ciências e Tecnologia, FCT, Universidade Nova de Lisboa

2829-516 Caparica, Portugal

{maa18899, pmrp, jf.martins}@fct.unl.pt

Abstract. This work describes the development of a system designed for renewable power generation integration. It continuously acquires wind, solar and temperature data, which is automatically correlated with energy parameters, obtained from renewable energy systems. The developed system was installed in an urban building equipped with photovoltaic cells and wind renewable generation. To validate the developed application, it was analyzed data of a wind generator and a set of photovoltaic panels, installed near to the weather station. The developed application allows, in addition to the acquisition of weather and energy data, their continuous monitoring and correlation through a graphical user interface, providing a friendly interactivity with the user.

Keywords: Weather Conditions, Energy Efficient Buildings, Renewable Energy Production.

1 Introduction

The use of fossil fuels - coal, oil and gas - as the primary energy source, was one of the main factors that made possible the rapid Humanity growth in the last century. However, from the oil crisis in the 70's, renewable energies started to have a significant role as a potential alternative to the fossil fuels [1]. The demand and energy consumption in any country is directly connected to its demography and development, having a strong impact on economic growth due to energy prices, particularly prices of fossil fuels [2].

Renewable electric energy production's efficiency depends not only on the existent technology, but also mainly on the weather conditions at a given location. Knowing the weather conditions at a specific place, it is possible to optimize systems that take advantage of renewable energy sources. Moreover, accurate meteorological records are crucial to accurately predict and monitor any energy production system. There are currently a wide range of commercial systems that are able to acquire weather conditions. From basic weather stations, capable of acquiring temperature, humidity and barometric pressure, to the more advanced, which, in addition, are able to measure, and record, wind speed and direction, precipitation and solar radiation. However, these commercial systems only allow the monitoring, and recording, of weather conditions.

In [3], a non-commercial weather acquisition system for monitoring a photovoltaic (PV) system is proposed. The system is able to acquire solar radiation and temperature, as well as the PV electric parameters, through a set of sensors connected to a microprocessor, used as the core of the system. The collected data are sent to a computer that works mainly as a database, and the system does not offer a graphical interface. The hardware involved makes the system considerably complex, thus not flexible to changes, making it difficult to adapt to other type of sensors. A similar system is also proposed in [4]. Here, authors stated as major advantages the hardware design and the possibility of operation in remote places, since the system is battery powered. In [5] is described a LabVIEW based system. All data is acquired and processed by LabVIEW, and the user can interact with the system through some graphical interface menus. However, the tool is only dedicated to PV systems. The system proposed in [6] proposes, as main characteristic, the potential to integrate different renewable energy sources, namely PV and wind generator systems. The core of the system is also a tool developed in LabVIEW. However, the graphical interface and the options presented are limited.

A full system for weather monitoring and renewable energy production correlation could become quite expensive, particularly when small renewable energy systems (< 5kWp) are considered. The system costs are the main reason why typically they are only used to: (i) demonstrate that the PV system is a reliable energy source for the given application, or (ii) to develop criteria for design and operation that optimize a PV system for its site and task [7].

In this work, we propose a tool for the management and management of weather data, as well as the correlation among meteorological of energy production data. Meteorological data will be acquired by a weather station equipped with several sensors; all of them installed in the same location as the renewable energy systems. There will be a periodic acquiring of the relevant atmospheric values, including wind speed and direction, temperature and solar radiation. Electrical power generation data will be obtained from another system designed for this purpose and already implemented [8]. With this work, it is our ambition to overcome some of the limitation of the existent systems. The developed tool has three main goals: (i) communication and data acquisition through the weather station; (ii) data processing and (iii) analysis and correlation of meteorological data with data acquired from renewable power generation systems. The developed tool was implemented in Matlab.

2 Contribution for Sustainability

Due to the continuous increase in energy consumption, connected to the fact that fossil fuels are becoming more rare and more expensive over the years, the market of renewable energy production systems has increased, mainly in the residential and small-medium enterprises sectors. The developed system, here presented, has two major goals: (1) to offer the possibility of estimating the energy production of new systems to be installed, and (2) to monitoring an existent renewable energy production system, correlating weather and energy data. This system aims to support and validate new technologies applied to renewable energy production systems, since “green” sources play an important role in today’s electric grid sustainability.

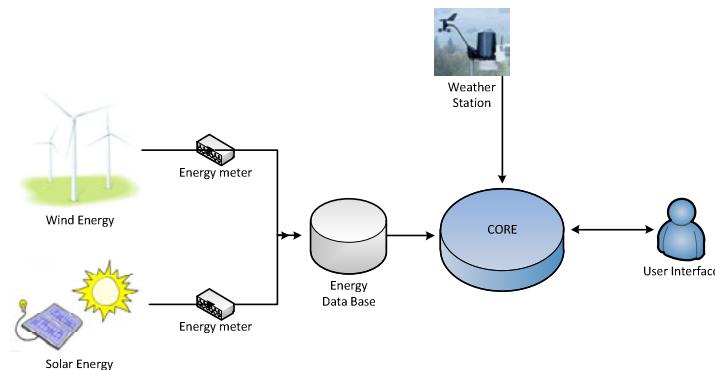


Fig. 1. Conceptual model

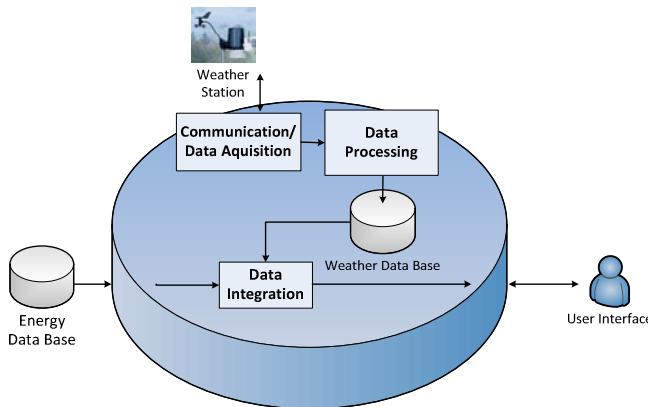


Fig. 2. Core block tasks

The developed system integrates two major areas of interest: (1) meteorological data and renewable electricity generation. Figure 1 presents the framework of the systems' conceptual model.

In Fig. 1, the "core" block is responsible for connecting the energy database with the weather station and the user. The energy database contains data daily acquired from energy meters connected to the renewable energy production systems – wind and photovoltaic generators. This block acts as the system brain, being responsible for acquiring the meteorological data. Moreover, it correlates the data acquired from the weather station with data acquired from the energy database, as presented in Fig. 2. The system offers a friendly user interface that allows a substantial control on parameterization and visualization data.

3 Monitoring and Correlation

The methodology adopted in this work aims to make possible the integration of future add-ons. To accomplish this purpose, the core block is divided in sub-blocks that

represent essential functions of the system. Those sub-blocks, presented in Fig. 3, will be briefly presented.

Communication/data acquisition

This block is responsible for the communication with the system hardware (*datalogger*). Once established the communication, it is possible to configure, in the weather station, several options, such as the reading interval, configure data or activate output ports. Furthermore, the nature of the data acquisition can be also configurated: the user can choose either to acquire data stored in the weather station memory or acquire real time data.

Data processing

In the data processing block, the data collected by the aforementioned routines is decoded. The processed data is then exported to MS Excel files (one file per day), which are stored in an internal database of the computer where the application is running.

Data integration

The renewable energy production is monitored by a system [8] that keeps daily records of wind generator and PV system energy data. This data integration block, depending on the available data in the databases, allows the user to choose the day and/or the date range that he desires to view/correlate.

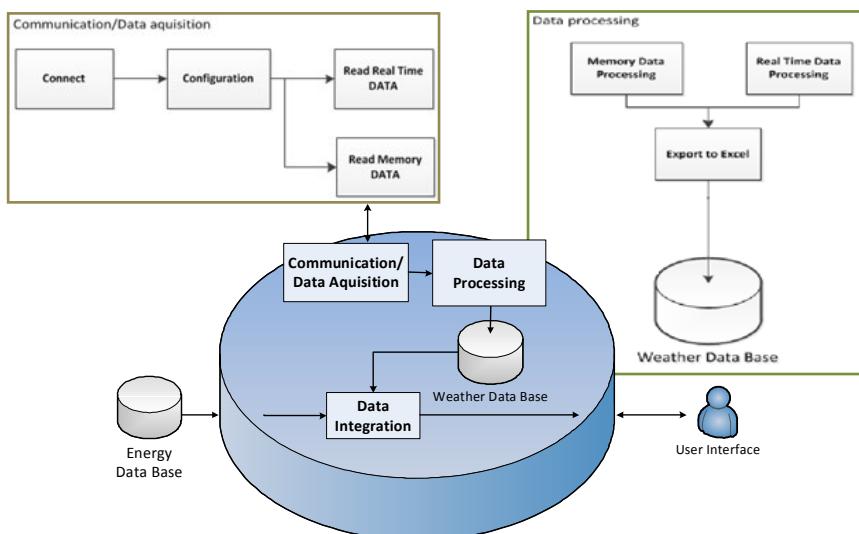


Fig. 3. “Core” sub-blocks

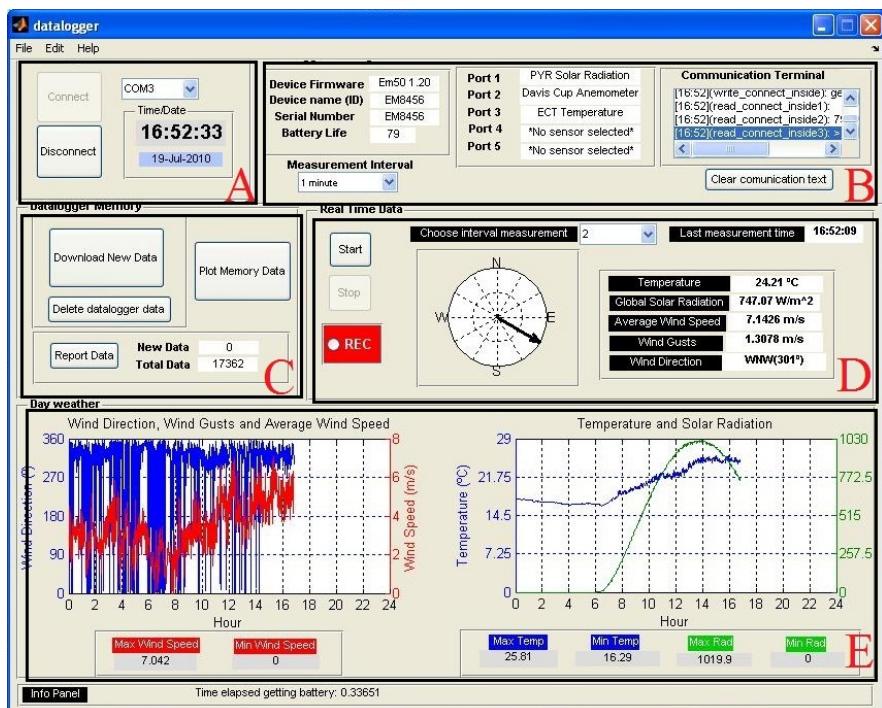
Table 1 presents the several parameters that are suitable for display and correlation. The developed system allows correlation analysis of these parameters (in two different graphics), and also presents a compass rose regarding the set of days selected, correlating wind speed and direction.

Table 1. List of parameters available for visualization and/or correlation

Electrical Energy Production Parameters (wind and PV)	Weather Parameters
Voltage (V)	Wind direction (°)
Current (A)	Wind Gusts (m/s)
Power Factor	Average Wind Speed (m/s)
Power (VA)	Temperature (°C)
Active Power (W)	Solar Radiation (W/m ²)
Reactive Power (VAr)	

User Interface

The user interface is one of the key points of the developed application, providing a friendly and appealing environment to the user. The main idea was to develop an interface that allows the user to have total control over all the involved parameters, either regarding hardware configuration or data acquisition. Figure 4 presents the main window of the implemented tool.

**Fig. 4.** Main window

In Fig. 4, the main interface is divided into several areas, each one related to a specific task. A summary of each block, identified with capital letters, is presented below.

- **A - Datalogger Control** – Starts and stops the connection with the weather station datalogger and gives information about current date and time;
- **B - Datalogger Settings** – Gives information about the weather station parameters: firmware version number of the device (Device Firmware), device name (Divide Name (ID)), unique serial number of the device (Serial Number), battery current status (Battery Life) and sensors assigned to each communication port. In the *Communication terminal* is possible to view all messages exchanged between the application and the equipment. It is also possible to define the data recording interval in the datalogger's internal memory (Measurement Interval);
- **C - Datalogger Memory** - Download of new data from datalogger's internal memory for the database in MS Excel files. It is possible to see the memory status, to clear the data logger internal memory and to open a new window for data correlation;
- **D – Real Time Data** - Shows weather conditions in real time with a minimum sample acquisition time of 2 seconds. The compass wind is updated in real time (depending on the user-defined acquisition time; 1 sec, 10 sec, ..., 1 min), indicating wind direction. Enables the record of meteorological data in real time to the database (Weather Data Base);
- **E – Day Weather** - Presents the weather conditions for the current day in two graphics. Also gives information about the maximum and minimum values achieved in the present day.

In addition to the main window, the user can open another window where he can choose the variables that he wants to correlate, as presented in Fig. 5-(A). The variables that can be correlated were previously presented in Table 1.

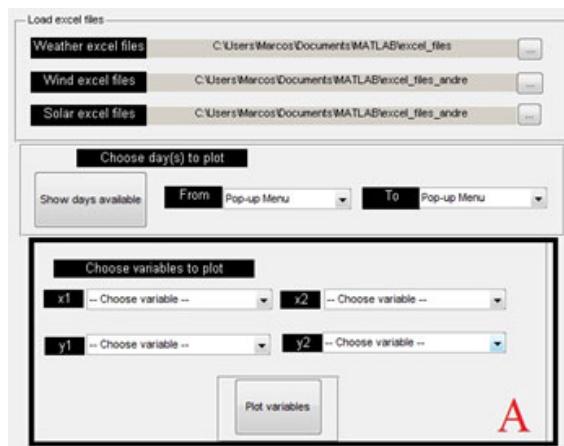


Fig. 5. Correlation window

In Fig. 5 window the user can create correlation graphics for the chosen variables, relating weather conditions and electrical parameters of the two renewable energy production systems. In addition to those variables, the total electrical energy produced

and the power factor, for the selected time range, are shown. A new window (Fig. 6) presenting the obtained correlation graphics is generated. This window also offers the user the possibility of applying multiple filters to the correlation graphics. The maximum and minimum values of each curve are also displayed.

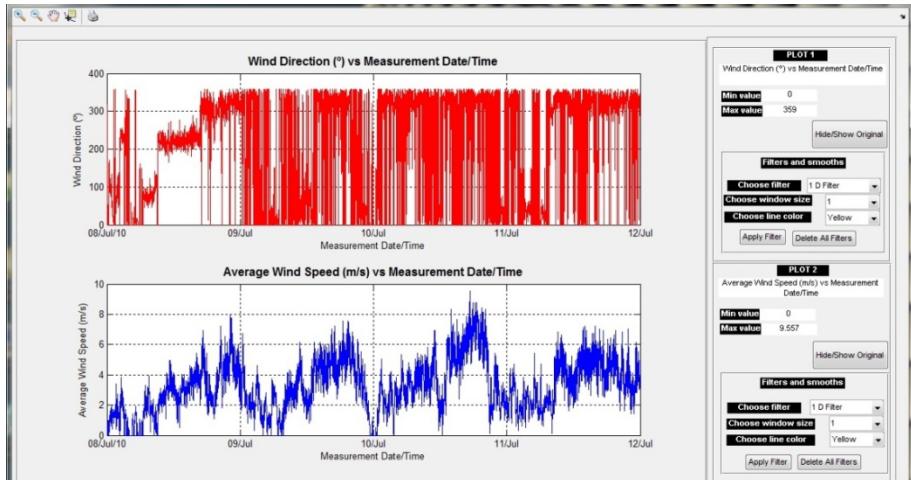


Fig. 6. Correlation window

4 Experimental Results

In this Section, experimental data obtained from the weather station and energy monitoring system, in the period of 23-29 of July 2010, is graphically presented. The weather station was set up acquisition sample intervals of one minute, resulting in 1440 readings per day. The energy monitoring system has acquisition sample intervals of 10 seconds, resulting in approximately 8000 readings per day. The wind generator is a 2 kW production system, and the PV has 460 Wp installed.

Correlation between weather conditions and wind generator production

Figure 7 shows three graphics were it is possible to see the wind speed, wind power, wind generator electrical power and the system power coefficient (C_p) over the period in analysis (Please note that the third graphic refers to the highlighted area in the second graphic). The highly oscillatory behaviour of the wind power and power coefficient are due to the sudden and unpredictable wind changes, which can be seen in the first graphic. The black line denotes average values of wind speed and C_p . Wind power and generator electrical power have identical behaviour, which means that the energy produced follows the wind speed variation. As expected, wind power is higher than generator electrical power, denoting a typical C_p of 0.3. The C_p is usually used as an indicator of the wind turbine efficiency and takes values between 0.3 and 0.4, for common real systems. According to Betz's law, only less than 59% (0.6) of the kinetic energy of wind can be converted into mechanical energy.

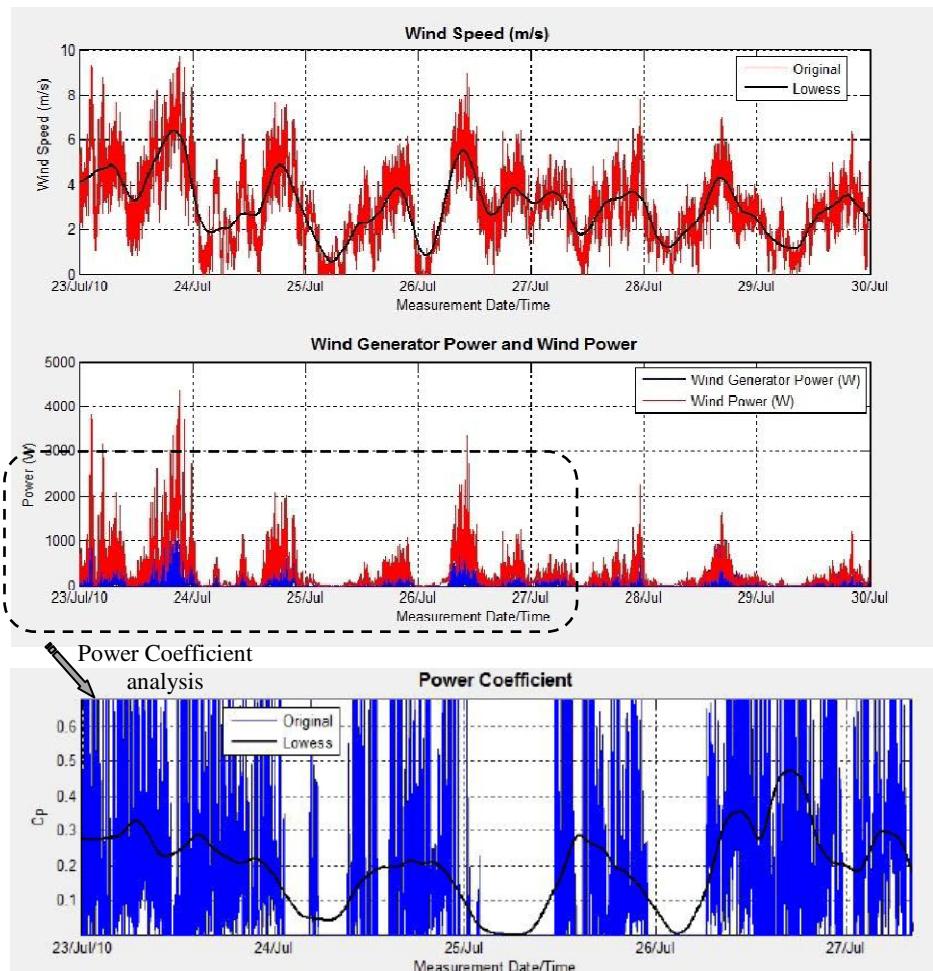


Fig. 7. Weather condition and wind generator energy production

Correlation between weather conditions and PV production

Fig. 8 shows the solar radiation, PV power and PV system efficiency. The electrical power generated by the PV system has, as expected, a similar behaviour as the solar radiation curve. The peak power reached is in concordance with the values of higher solar radiation measured on each day.

Another key point about this PV system is its efficiency. PV cells specified efficiency is obtained under Standard Test Conditions (STC), which means that the real PV efficiency largely depends on the place, and respective weather conditions, where the system is installed. For the system under analysis, the STC efficiency is about 7.2%. However, as it is possible to see in Fig. 8, the real efficiency of the system is around 5%. For PV panels the efficiency increases with solar radiation but decreases with the temperature of the panel. As expected, maximum efficiency occurred when

maximum solar radiation was achieved. But, for maximum solar radiation, a high temperature value was also measured. High temperatures, over 25-27°C, influences negatively the PV efficiency. This is the reason why, in Fig. 8, the efficiency of the system was slightly higher in point 2 than in 1, even with less radiation the temperature was lower in point 2.

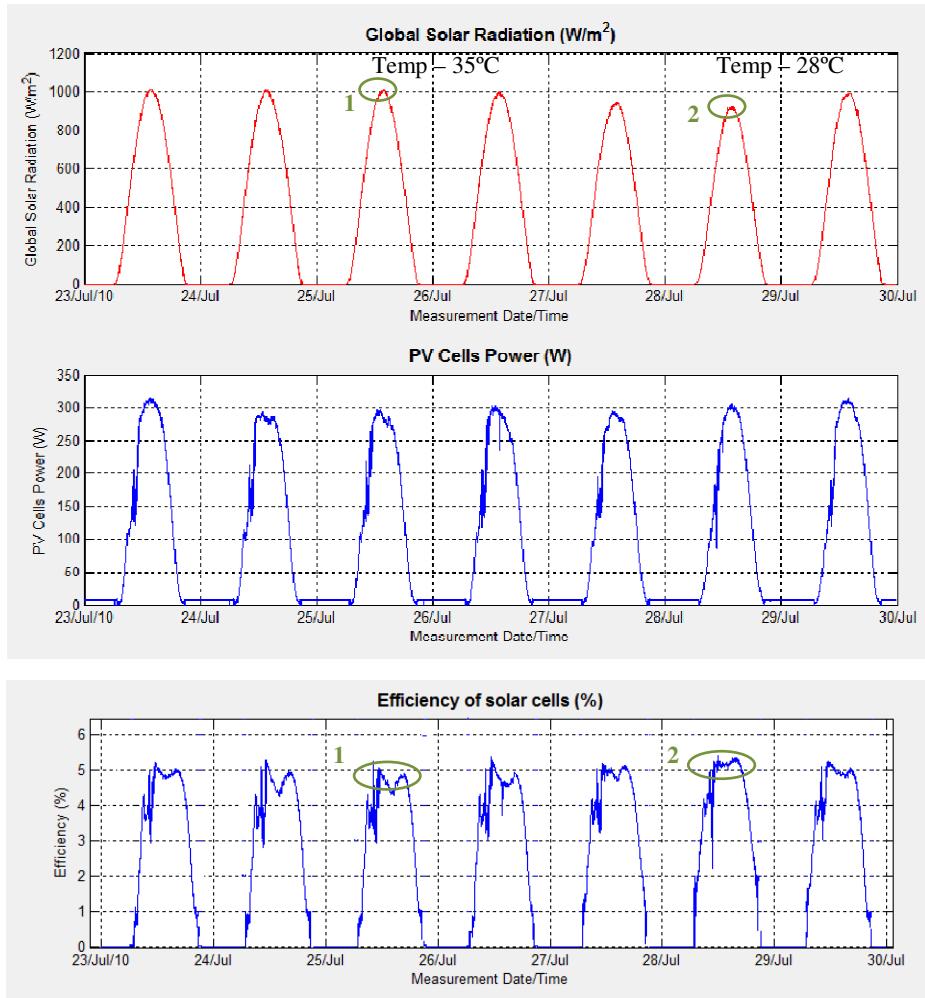


Fig. 8. Weather condition and PV system

5 Conclusions

This work has presented a system that integrates weather conditions information and energy parameters from renewable energy power generation system. The system incorporates an integrated analysis of the multiple variables involved, through an

interactive graphical user interface. One of the advantages of an accessible and interactive user interface is the simple analysis of the renewable power systems generated electrical energy in relation to the existing weather conditions.

To validate the developed application, it was used data from a wind generator and photovoltaic panels, installed near to the weather station. The application developed in this work allows, in addition to the acquisition of weather and energy data, their monitoring and correlation through a simple and attractive graphical user interface, providing an easy interactivity with the user.

The implemented system can be used by institutions or companies related to renewable energy systems, meteorology, or even by the private user who is interested in monitoring the production of electricity from renewable energy systems installed in their homes.

Acknowledgments. This work was supported by FCT (CTS multiannual funding) through the PIDDAC Program funds.

References

1. Elhadidy, M., Shaahid, S.: Parametric Study of Hybrid (wind+solar+diesel) power generating Systems. *Renew Energy* 21(2), 129–139 (2000)
2. IEA: World Energy Outlook 2009. International Energy Agency Publications (2008)
3. Benghanem, M., Maafi, A.: Data Acquisition System for Photovoltaic Systems Performance Monitoring. *IEEE Transactions on Instrumentation and Measurement* 47, 30–33 (1998)
4. Mukaro, R., Carelse, X.F.: A Microcontroller-Based Data Acquisition System for Solar Radiation and Environmental Monitoring. *IEEE Transactions on Instrumentation and Measurement* 48, 1232–1238 (1999)
5. Forero, N., Hernández, J., Gordillo, G.: Development of a monitoring system for a PV solar plant. *Energy Conversion and Management* 47(15-16), 2329–2336 (2006)
6. Koutroulis, E., Kalaitzakis, K.: Development of an integrated data-acquisition system for renewable energy sources systems monitoring. *Renewable Energy* 28(1), 139–152 (2003)
7. Blaesser, G.: PV system measurements and monitoring: The European experience. *Solar Energy Materials and Solar Cells* 47(1-4), 167–176 (1997)
8. Jorge, A., Guerreiro, J., Pereira, P., Martins, J., Gomes, L.: Energy Consumption Monitoring System for Large Complexes. In: Camarinha-Matos, L.M., Pereira, P., Ribeiro, L. (eds.) DoCEIS 2010. IFIP AICT, vol. 314, pp. 22–24. Springer, Heidelberg (2010)

Comparison of Different Modulation Strategies Applied to PMSM Drives Under Inverter Fault Conditions

Jorge O. Estima and A.J. Marques Cardoso

University of Coimbra, FCTUC/IT,
Department of Electrical and Computer Engineering,
Pólo II – Pinhal de Marrocos, P – 3030-290, Coimbra, Portugal
Phone/Fax: +351 239 796 232/247
jestima@ieee.org, ajmcardoso@ieee.org

Abstract. This paper presents a comparative study of two distinct modulation strategies applied to a permanent magnet synchronous motor drive, under inverter faulty operating conditions. A rotor field oriented control is used and a performance evaluation is conducted by comparing a hysteresis current control with a space vector modulation technique. Three different operating conditions are considered: normal situation, a single power switch open-circuit fault and a single-phase open-circuit fault. In order to compare the drive performance under these conditions, global results are presented concerning the analysis of some key parameters such as motor efficiency, power factor, electromagnetic torque and currents rms and distortion values.

Keywords: Permanent magnet synchronous motor, hysteresis current control, space vector modulation, inverter open-circuit faults.

1 Introduction

Permanent magnet synchronous motors (PMSM) employed in AC drive systems are usually fed by six-switch three-phase voltage source inverters. Due to their complexity, it is known that these devices are very susceptible to the occurrence of failures, either in the power stage or in the control subsystem. These can be broadly classified as short-circuit faults and open-circuit faults. Short-circuits represent the most serious class of faults. In case of a single power device short-circuit, the second switch in the same inverter leg has to be turned off immediately to avoid a dangerous shoot-through inverter failure. Open-circuit failures may occur when, for any reason, the semiconductor is disconnected, damaged or due to a problem in the gate control signal.

Some studies can be found in the literature concerning the analysis of PMSM drives under different faulty operating conditions. In [1], the effects on currents, voltages and torque of different drive failures such as switch-on failures, single-phase open-circuit, switch-off and DC supply failures, are investigated. The steady state and dynamic response of a PMSM drive to a single-phase open-circuit fault is investigated in [2], and in [3] symmetrical and asymmetrical short-circuit faults are considered.

Major potential faults that can occur in PMSM drives are discussed and simulated in [4]. Single-phase open-circuit and short-circuit faults, three-phase short-circuit faults and single switch-on failures are considered.

Several performance analysis of a PMSM drive under a single power switch and single-phase open-circuit faults in the inverter are reported in [5]-[7]. Through the evaluation of some key parameters, it was concluded that a single-phase open-circuit fault has a greater negative impact on the PMSM performance than a single power switch open-circuit failure.

This paper intends to present a comparative analysis between two different PWM strategies applied to PMSMs control, namely, a rotor field oriented control with hysteresis current controllers and with space vector PWM (SV-PWM), under inverter faulty operating conditions. Three distinct operating conditions are considered – normal behavior, a single power switch open-circuit fault and a single-phase open-circuit fault in the inverter. A typical three-phase diode bridge rectifier, a VSI and a PMSM are used, as shown in Fig. 1.

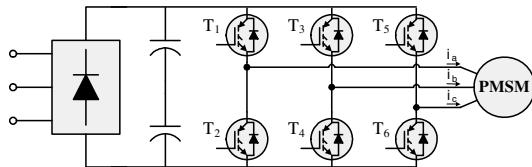


Fig. 1. Structure of the PMSM drive

The PMSM performance evaluation for both control strategies is based on the analysis of some key parameters such as motor efficiency, power factor, electromagnetic torque and currents rms and harmonic distortion values.

2 Contribution to Technological Innovation for Sustainability

Due to the arising concerns about global warming and energy resources constraints, there is nowadays an increasing demand for high-efficient energy conversion systems. Therefore, and considering also the large impact of electric motor drives energy consumption in the worldwide industry, the subject discussed in this work can contribute to the technological innovation for sustainability since, among other things, a comparison of efficiency levels is performed.

3 PMSM Dynamic Model Equations

Typical PMSM mathematical models found in the literature do not take iron losses into account. For this reason, in order to obtain a more accurate modeling, especially for the iron losses, a dedicated parameter has been considered aimed at accounting for the iron losses in the stator core, specifically the eddy current losses. These are modeled by a resistor R_c which is inserted in parallel with the magnetizing branch, so that the power losses depend on the air-gap flux linkage [6]. Therefore, assuming that the saturation is neglected, the electromotive force is sinusoidal and a cageless rotor, the stator dq equations in the rotor reference frame are:

$$v_d = R_s i_d + L_d \frac{di_{md}}{dt} - \omega L_q i_{mq} \quad (1)$$

$$v_q = R_s i_q + L_q \frac{di_{mq}}{dt} + \omega L_d i_{md} + \omega \psi_{PM} \quad (2)$$

where v_d and v_q are the dq axes voltage components, R_s the stator winding resistance, i_d and i_q the dq axes supply currents, L_d and L_q the dq axes inductances, i_{md} and i_{mq} the dq axes magnetizing currents, ω the fundamental frequency and ψ_{PM} the flux linkage due to the rotor magnets.

The PMSM electromagnetic torque T_e equation is given by:

$$T_e = \frac{3}{2} p [\psi_{PM} i_{mq} + (L_d - L_q) i_{md} i_{mq}] \quad (3)$$

being p the machine pole pairs number.

4 Results

The modeling and simulation of the drive was carried out using the Matlab/Simulink environment, in association with the Power System Blockset software toolbox. A rotor field oriented control strategy was implemented for a PMSM employing a hysteresis current control, in the abc reference frame, and a space vector PWM (SV-PWM). The value of the hysteresis band was defined to 0.3 A and the SV-PWM switching frequency was chosen to be 8 kHz. Three different operating conditions are considered: normal situation, an inverter single power switch open-circuit fault (IGBT T1) and a single-phase open-circuit failure (double fault in IGBTs T1 and T2).

The PMSM phase currents are analyzed by the calculation of their rms and distortion values using the Total Waveform Distortion (TWD) defined as:

$$\text{TWD} = \frac{\sqrt{X_{\text{rms}}^2 - X_1^2}}{X_1} \times 100\% \quad (4)$$

being X_{rms} the waveform rms value and X_1 its respective fundamental component.

In order to study the electromagnetic torque developed by the PMSM for the considered cases, a Total Waveform Oscillation (TWO) parameter is also introduced, which is given by:

$$\text{TWO} = \frac{\sqrt{T_{e_{\text{rms}}}^2 - T_{e_{\text{dc}}}^2}}{|T_{e_{\text{dc}}}|} \times 100\% \quad (5)$$

where $T_{e_{\text{rms}}}$ and $T_{e_{\text{dc}}}$ stand for the electromagnetic torque rms and average values, respectively.

Finally, for all the considered operating conditions, a constant load torque equivalent to 28% of the PMSM rated torque is assumed, together with a reference speed of 1200 revolutions per minute.

4.1 Normal Operating Conditions

Fig. 2 presents the time-domain waveforms of the motor phase currents obtained for a hysteresis current control and for a space vector PWM technique under normal operating conditions. Fig. 3 presents their corresponding rms and distortion values.

As expected, under healthy operating conditions, the PMSM phase currents are practically sinusoidal, containing a well-defined fundamental component and low amplitude high-frequency noise. However, in spite of their rms values are the same for both modulation strategies (Fig. 3(a)), it can be seen that with a SV-PWM technique, it is possible to achieve lower distortion values (Fig. 3(b)).

Fig. 4(a) and Fig. 4(b) present the spectrograms of the electromagnetic torque and their corresponding time-domain waveforms for a hysteresis current control and for a SV-PWM, respectively. Under normal operating conditions, there are no appreciable differences between both cases, leading to a similar TWO value.

Regarding the PMSM power factor, the results in Fig. 5(a) show that with a hysteresis current control, the obtained power factor value is considerably lower than with a SV-PWM strategy. This is justified by the larger rms values of the supplying voltage generated by the hysteresis controllers, which lead to the increasing of the apparent power and the subsequent decrease of the machine power factor.

Fig. 5(b) presents the PMSM efficiency values for both the considered modulation strategies. Once more, the use of SV-PWM allows to achieve higher efficiency values since, as previously mentioned, with a hysteresis current control, larger voltage rms values are applied to the machine, contributing considerably to the increase of the PMSM iron losses.

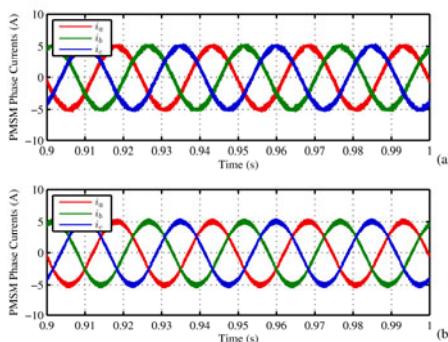


Fig. 2. Time-domain waveforms of the PMSM phase currents under normal operating conditions: (a) hysteresis current control; (b) SV-PWM

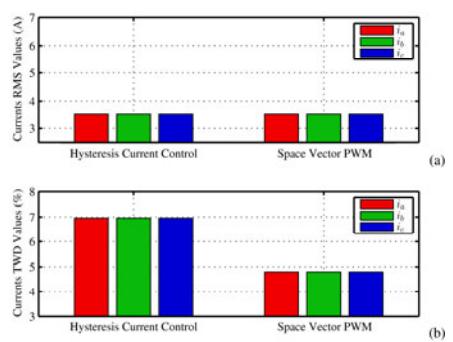


Fig. 3. PMSM phase currents rms (a) and TWD values (b) under normal operating conditions

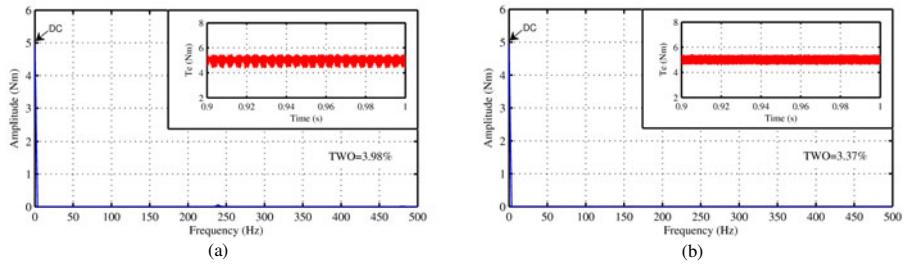


Fig. 4. Spectrograms of the PMSM electromagnetic torque and its corresponding time-domain waveforms under normal operating conditions: (a) hysteresis current control; (b) SV-PWM

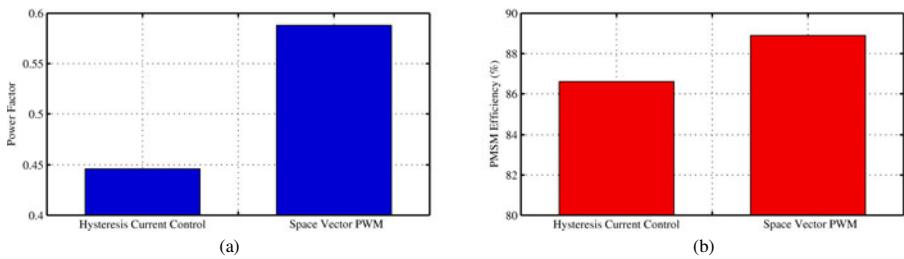


Fig. 5. PMSM power factor and efficiency results under normal operating conditions: (a) power factor values; (b) efficiency values

4.2 Single Power Switch Open-Circuit Fault

Fig. 6 presents the time-domain waveforms of the motor phase currents obtained for a hysteresis current control and for a space vector PWM technique with a single power switch open-circuit fault in transistor T1. Fig. 7 presents their corresponding rms and distortion values. It can be clearly seen that, under these conditions, the motor phase currents do not have a sinusoidal shape anymore. This unbalanced inverter topology also influences the currents rms values, where the affected phase (phase *a*) will have the lowest value, increasing the remaining values of the healthy inverter legs. Comparing with normal operating conditions, the TWD values increase significantly, particularly for the faulty phase. However, comparing both modulation strategies, the SV-PWM has a better behavior since it generates less harmonic distortion on the two healthy phases.

The results presented in Fig. 8 show that the PMSM electromagnetic torque is no longer constant, containing harmonics multiple of the fundamental currents frequency. The main pulsating component is less significant for the SV-PWM control, which contributes to a lower TWO value and a smoother torque.

Fig. 9 presents the results concerning the PMSM power factor and efficiency. Comparing with the healthy operating conditions, although all values are negatively affected by the fault, the SV-PWM technique has a better performance since it allows to achieve higher power factor and efficiency values.

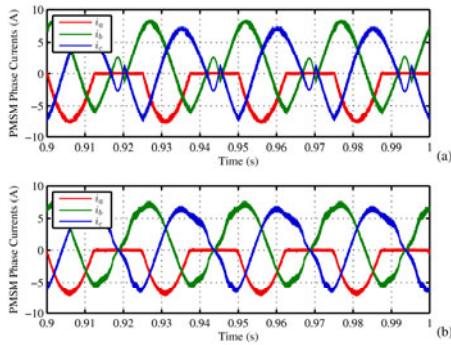


Fig. 6. Time-domain waveforms of the PMSM phase currents for an open-circuit fault in transistor T1: (a) hysteresis current control; (b) SV-PWM

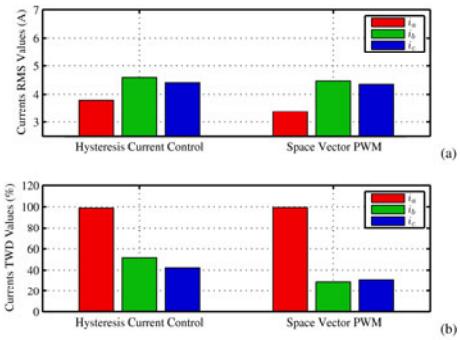


Fig. 7. PMSM phase currents rms (a) and TWD values (b) for an open-circuit fault in transistor T1

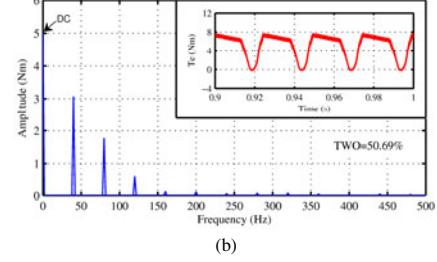
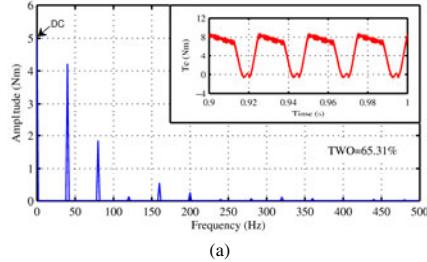


Fig. 8. Spectrograms of the PMSM electromagnetic torque and its corresponding time-domain waveforms for an open-circuit fault in transistor T1: (a) hysteresis current control; (b) SV-PWM

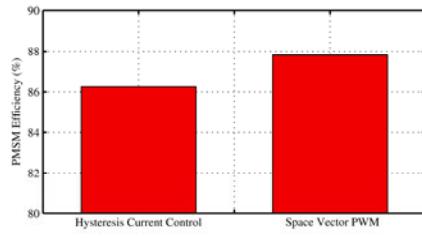
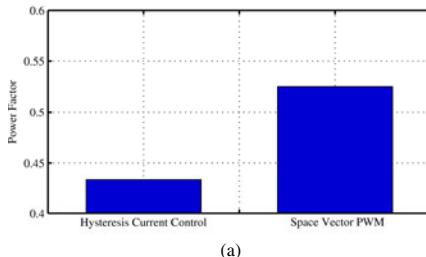


Fig. 9. PMSM power factor and efficiency results for an open-circuit fault in transistor T1: (a) power factor values; (b) efficiency values

4.3 Single Phase Open-Circuit Fault

Fig. 10 and Fig. 11 present the time-domain waveforms of the motor phase currents and their corresponding rms and distortion values obtained for the two considered techniques with a single-phase open-circuit fault in phase a . As expected, under these conditions the current in phase a is null. However, it is verified that the two remaining currents amplitude is larger for a hysteresis current control, which contributes to a greater thermal stress imposed on the stator windings insulation. Furthermore, the TWD results show that the SV-PWM has a better behavior since it generates much less harmonic distortion than the hysteresis current control.

The time-domain waveforms of the electromagnetic torque developed by the PMSM (Fig. 12) confirm its very pulsating nature. Comparing both modulation strategies, the SV-PWM leads to the creation of a less pulsating torque, which means that as far as the motor is concerned, the produced mechanical stresses are reduced, comparing with a hysteresis current control.

Regarding the power factor results presented in Fig. 13(a), despite the noticeable increase of the PMSM power factor with a hysteresis current control, a higher value is obtained with a SV-PWM technique.

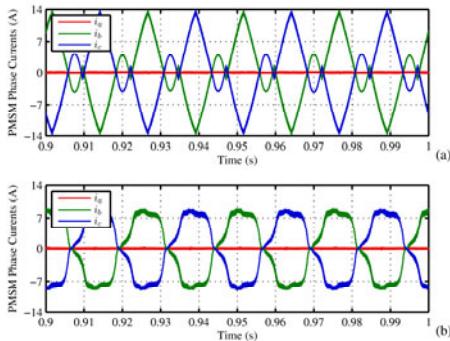


Fig. 10. Time-domain waveforms of the PMSM phase currents for a single-phase open-circuit fault in phase a : (a) hysteresis current control; (b) SV-PWM

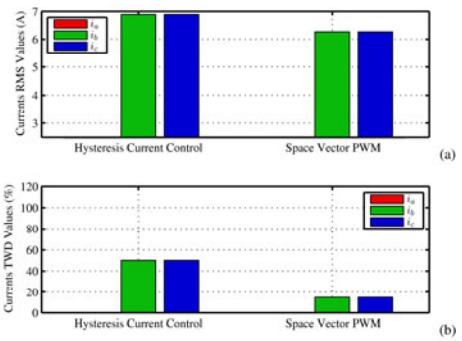


Fig. 11. PMSM phase currents rms (a) and TWD values (b) for a single-phase open-circuit fault in phase a

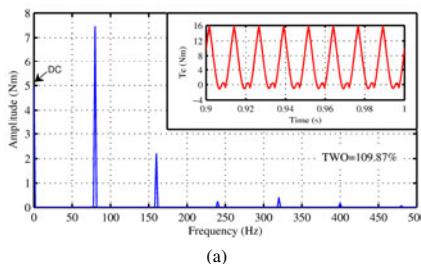


Fig. 12. Spectrograms of the PMSM electromagnetic torque and its corresponding time-domain waveforms for a single-phase open-circuit fault in phase a : (a) hysteresis current control; (b) SV-PWM

Fig. 13(b) presents the PMSM efficiency results for the two considered strategies. Due to the larger motor phase currents rms values, the efficiency values are severely affected by this fault type, when comparing to the normal operating conditions. Nevertheless, with a SV-PWM technique it is possible to obtain a higher efficiency value than with a hysteresis current control.

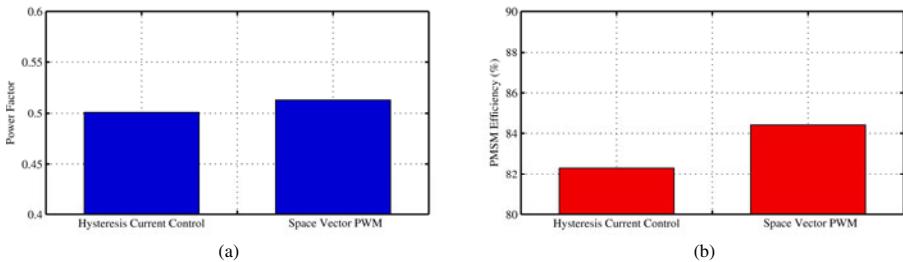


Fig. 13. PMSM power factor and efficiency results for a single-phase open-circuit fault in phase a : (a) power factor values; (b) efficiency values

5 Conclusions

The results presented in this paper allow to conclude that with a SV-PWM technique applied to a PMSM drive, it is possible to achieve a better performance than with a hysteresis current control. Under both normal and faulty operating conditions, by using the SV-PWM technique it is possible to obtain lower rms and distortion values in the motor phase currents, a less pulsating electromagnetic torque and higher power factor and efficiency values, as compared to the use of the hysteresis current control.

Acknowledgments. The authors gratefully acknowledge the financial support of the Portuguese Foundation for Science and Technology (FCT) under Project No. SFRH/BD/40286/2007 and Project No. PTDC/EEA-ELC/105282/2008.

References

1. Bianchi, N., Bolognani, S., Zigliotto, M.: Analysis of PM synchronous motor drive failures during flux weakening operation. In: 27th Annual IEEE Power Electronics Specialists Conference, Baveno, Italy, June 23-27, vol. 2, pp. 1542–1548 (1996)
2. Welchko, B.A., Jahns, T.M., Hiti, S.: IPM synchronous machine drive response to a single-phase open circuit fault. *IEEE Transactions on Power Electronics* 17(5), 764–771 (2002)
3. Welchko, B.A., Jahns, T.M., Soong, W.L., Nagashima, J.M.: IPM synchronous machine drive response to symmetrical and asymmetrical short circuit faults. *IEEE Transactions on Energy Conversion* 18(2), 291–298 (2003)
4. Sun, T., Lee, S.H., Hong, J.P.: Faults analysis and simulation for interior permanent magnet synchronous motor using Simulink@MATLAB. In: International Conference on Electrical Machines and Systems, Seul, South Korea, October 8-11, pp. 900–905 (2007)

5. Estima, J.O., Cardoso, A.J.M.: Performance evaluation of permanent magnet synchronous motor drives under inverter fault conditions. In: XVIII International Conference on Electrical Machines, Vilamoura, Portugal, CD-ROM, September 6-9, p. 6 (2008)
6. Estima, J.O., Cardoso, A.J.M.: Performance evaluation of DTC-SVM permanent magnet synchronous motor drives under inverter fault conditions. In: 35th Annual Conf. of IEEE Industrial Electronics Society, Porto, Portugal, November 3-5, p. 6 (2009)
7. Estima, J.O., Cardoso, A.J.M.: Impact of Inverter Faults in the Overall Performance of Permanent Magnet Synchronous Motor Drives. In: IEEE International Electric Machines and Drives Conference, Miami, USA, pp. 1319–1325 (May 2009)

Optimization of Losses in Permanent Magnet Synchronous Motors for Electric Vehicle Application

Ana Isabel León-Sánchez, Enrique Romero-Cadaval,
María Isabel Milanés-Montero, and Javier Gallardo-Lozano

Research, Development and Innovation group *Power Electrical & Electronic Systems PE&ES*, Escuela de Ingenierías Industriales, Universidad de Extremadura,
Avda. de Elvas s/n, 06006 Badajoz, Spain
{aleon@peandes, eromero@, milanes@, jagallardo@peandes}unex.es

Abstract. The aim of this paper is to analyze the influence of some parameters related with the permanent magnet motor control, which are being considered as an attractive alternative for electrical vehicle application where it is important to minimize the losses to increase the vehicle autonomy. In this paper, the attention is focused on the selection of the speed controller parameters and its impact both in mechanical losses as in the corresponding torque-speed trajectories. Moreover, the dependency of the switching frequency over the total losses in the motor-converter device is shown. The analysis determines the optimum value for the switching frequency and the results are compared with commercial servos.

Keywords: Electrical Vehicle, Permanent Magnet Synchronous Motor, Optimized losses.

1 Introduction

The growing interest in electric vehicles comes from the early 1990's, driven by rising fuel prices, high economic dependence between nations and strong climate impact due to the several pollutant emissions from traditional transport.

Currently, the vehicles are being developed with propulsion by energy of easy distribution and from different sources, such as electric power, not forgetting its easy availability in urban areas. Other advantage of using electric motors is the reduction of noise pollution. The proliferation of renewable energy for electricity generation helps the development of transports with electric motors.

With the increase in the use of AC motors compared with the DC motors, due to its lower cost and maintenance [1], and the mechanical benefits that they offer, new techniques have been developed for analysis. The Permanent Magnet Synchronous Motor, PMSM, does not require an external power source for excitation and exhibit high efficiency ratios compares with induction motors that have high losses in the short-circuit rotor. The application of digital control since current, torque and flux of these motors, leads the need of a compact and reliable model. These models must incorporate the essential elements of electromagnetic and mechanical behavior, for both the transient and the steady-state. The new analysis techniques of control motors places the PMSM as the motor selected for use in high performance electric vehicles achieved through the high ratio of torque-loading and adequate dynamic capacity [2].

2 Contribution to Sustainability

The use of electrical vehicles, at least in an urban environment, will contribute to the sustainability of the transport systems, decreasing the dependence of fossil fuels, decreasing also the emission of CO₂. Also, the energy used for this kind of vehicles could be produced by renewable energy generation systems.

A key factor for electrical vehicles use promoting is their autonomy. In this paper, the influence of the coefficients used for the PI speed controllers on the overall motor response is important for minimizing the losses in vehicle applications, and so to increase vehicle autonomy, is studied.

3 PMSM Model

The behavior of a PMSM could be modeled by equations [3]:

$$u_{sq} = R_s i_{sq} + \omega_e \phi_{sd} + \frac{d(L_{sq} i_{sq})}{dt}, \quad (1)$$

$$u_{sd} = R_s i_{sd} - \omega_e \phi_{sq} + \frac{d(L_{sd} i_{sd} + \phi_f)}{dt}, \quad (2)$$

$$t_e = \frac{3}{2} p ((L_{sd} - L_{sq}) i_{sd} i_{sq} + \phi_f i_{sq}), \quad (3)$$

$$\omega_m = \frac{\omega_e}{p}, \quad (4)$$

$$t_e - t_L = \omega_m F_r + J \frac{d\omega_m}{dt}. \quad (5)$$

These equations could be easily modeled in a simulation environment as, for example, MATLAB/SIMULINK. In this work, a MA-55 INFRANOR MAVILOR PMSM has been used for the simulation tests with the parameters¹ shown in Table 1.

Table 1. Variables and parameters used for modeling the PMSM

Symbol	Nomenclature	Unit	Values	General View
$u_{sq}; u_{sd}$	d-q Voltage components	V	-	
$i_{sd}; i_{sq}$	d-q Stator current components	A	-	
ω_e	Motor electrical speed	r.p.s.	-	
ϕ_f	Magnetic flux linkage	V s/rad	-	
t_e	Electromagnetic torque	N m	-	
ω_m	Rotor speed	r.p.s.	-	
t_L	Load torque	N m	-	
R_s	Stator resistance	Ω	0.7	
L_{sd}	d-axis inductance	H	$1.871 \cdot 10^{-3}$	
L_{sq}	q-axis inductance	H	$1.616 \cdot 10^{-3}$	
p	Pole pairs	Number	4	
J	Moment of inertia	kg m^2	$3.6 \cdot 10^{-3}$	
F_r	Viscous friction coefficient	$\text{N} \cdot \text{m} \cdot \text{s}/\text{rad}$	$2.25 \cdot 10^{-3}$	

¹ $R_s, L_{sd}, L_{sq}, p, J$ and F_r .

4 PMSM Control

The open loop control of a synchronous motor with variable frequency can develop a satisfactory variable speed when the motor works with stable values of the torque, without many requirements on speed. When the drive specifications require fast dynamic response and high accuracy in speed or torque control, the open loop control does not offer this possibility. That is why it is necessary to operate the motor in closed loop, where the operation dynamic drive system plays a fundamental role like an indicator of the system which takes part [4]. Control strategies can be classified in scalar control and vector control categories.

In scalar control the fed-voltage changes proportionally with the frequency, but this type of control is used only when motor works in a low speed range [5].

Vector control, (usually implemented with Digital Signal Processors, DSP), is used when required specifications are more exigent (related to speed or position). There are two principal techniques:

- I. Field oriented control (FOC). In this control, the stator current is controlled in bases of a *synchronous d-q frame* [2], (Fig. 1 (a)).
- II. Direct Torque Control (DTC). This control tries to achieve the desired torque by applying a vector voltage pattern table, (Fig. 1 (b) [6]).

Vector control usually uses an encoder to determine the rotor position. At present, research is focused on obtaining algorithms that estimate this rotor position without using encoders, (Sensorless Control [7], [8] and [9]).

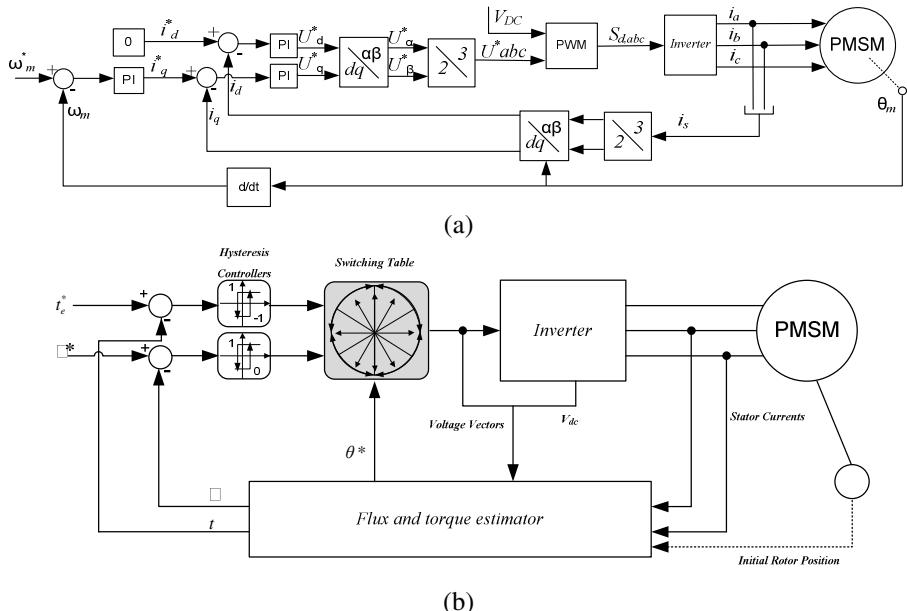


Fig. 1. Control System for PMSM: (a) FOC scheme, (b) DTC scheme

5 PMSM Performance

In this section it will be analyzed the performance of the modeled PMSM. This performance evaluation will be carried out by analyzing torque-speed curves when a pre-established torque-speed pattern is applied to the motor.

In this work it is used an in-wheel PMSM model controlled by FOC and based in the concept of active flux [10],[11], defined as the flux that multiplies current i_q in the expression (3). This model has been implemented using MATLAB/SIMULINK, (Fig. 2).

Two cases are studied to compare the influence of the PI coefficients of the angular speed regulator in the performance curves and also in the motor losses.

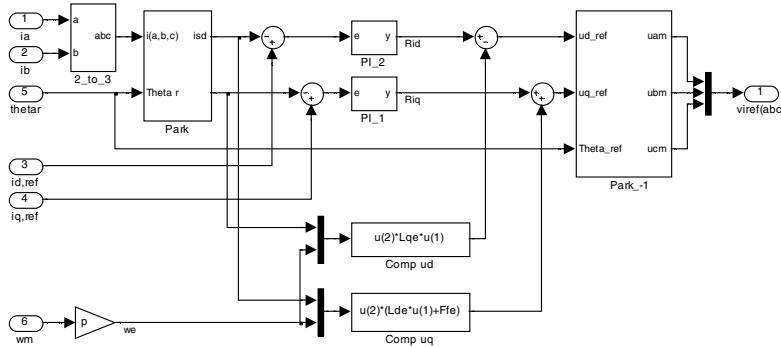


Fig. 2. Block diagram of the proposed control strategy for the PMSM in SIMULINK

The coefficients used for the PI speed controllers, which influence on the system will be studied by simulation, are listed in Table 2, while the coefficients used for the current controllers are listed in Table 3 (the same coefficients are used for both components d and q because the difference between L_{sd} and L_{sq} is negligible and these parameters are not under this paper scope).

The overall motor response, for cases A and B, is shown in Fig. 3. Detailed information of transient states, in a speed-torque plane, is represented in Fig. 4, where it can be observed how the motor reaches a final steady state from another steady state, after a change in speed or torque reference values takes place, and it could be observed if there are or not oscillations that will produce additional losses.

These graphs are employed to determine the energy used in the transient state when is produced a speed or torque variation, by using the integral expression (6).

Table 4 shows the energy needed for changing the motor steady state

$$\Delta E = \int_{t_0}^{t_1} (t_e - t_L) \omega dt. \quad (6)$$

Table 2. Speed controller coefficients

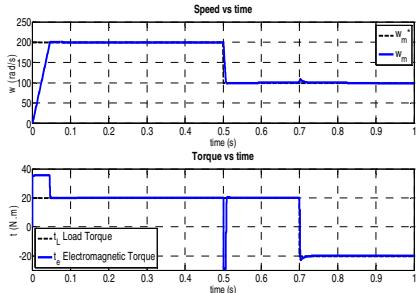
Symbol	Case A	Case B
k_p	5	5
k_i	500	5000

Table 3. Current controller coefficients

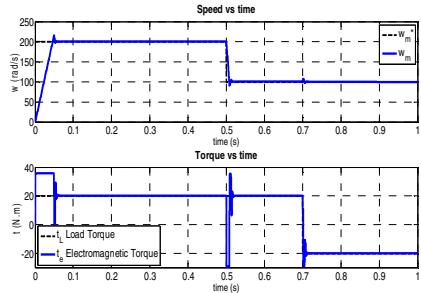
Symbol	Component d	Component q
k_p	50	50
k_i	4000	4000

Table 4. Energy (J)

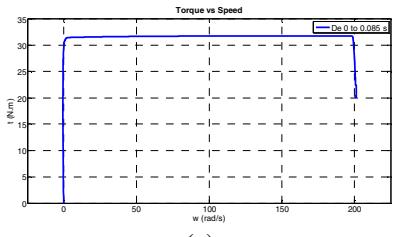
State	Case A	Case B	Relative variation (%)
Transient 1 (0 to 0.085)	197.4518	202.5904	2.5
Transient 2 (0.5 to 0.525)	-31.3359	-26.325	19.03
Transient 3 (0.7 to 0.74)	28,5789	29.0108	1.5
Total Profile	194,6948	205,2762	5.15



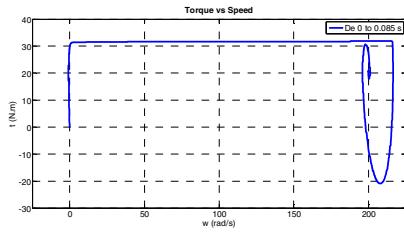
(a)



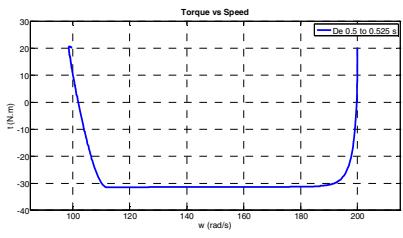
(b)

Fig. 3. Speed and torque standard curves (a) Case A, (b) Case B

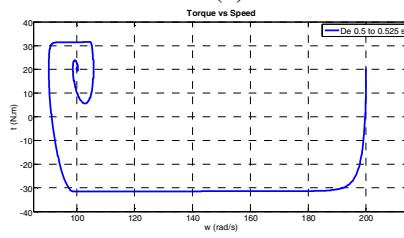
(a)



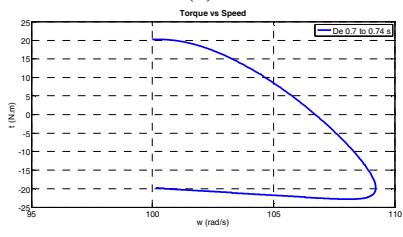
(b)



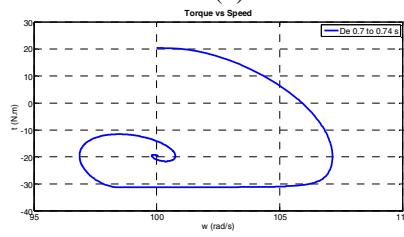
(c)



(d)



(e)



(f)

Fig. 4. Torque/Speed curves: Transient 1: (a) Case A, (b) Case B; Transient 2: (c) Case A, (d) Case B; Transient 3: (e) Case A, (f) Case B

Three different transient intervals are established. The losses determined by (6) depend highly on the PI coefficients selected for the speed controller in the Transient 2 interval, (achieving a nearly 20% of variation), because the torque and speed oscillation causes additional friction and electrical (mainly due to Joule effect) losses. In the other transients this dependence it is not so high, and it is due to the non-linearity of the motor and converter behavior.

6 Switching Frequency Influence

In this section, the losses that depend on the switching frequency, principally the converter losses, (conduction and switching), and the additional losses caused by the ripple current components produced in the motor windings will be analyzed.

For determining these losses a simulation test using PSIM has been done, (Fig. 5), because the inverter models calculate directly the conduction and switching losses. The PMSM has been modeled using the parameters of Table 1. The model has been simulated for different switching frequency values and for each simulation it has been wrote down the converter losses and then, the additional losses has been calculated. Every simulation is referred to a motor supplied by an ideal sinusoidal voltage that will produce sinusoidal currents calculating the losses by:

$$P_{ad} = P_{conv} - P_{sin} = 3R(I_{rms}^2 - I_1^2). \quad (7)$$

The results of the simulation set are drawn in Fig. 6. As it was expected, the converter losses increase with the switching frequency and the additional electrical motor losses decrease with it. The curves cross near 400 Hz, however, they show that the upslope of the losses of the inverter circuit does not grow in the same order of the decreasing load curve. Consequently, the frequency where the total losses are minimized in the simulation set is moved to frequencies higher than the corresponding to the point of intersection, (between the total losses of inverter circuit and load losses). If one considers optimizing the total losses, sum of converter and additional motor losses, the Fig. 6 shows that for this motor the optimum switching frequency is near 1 kHz (1350 Hz).

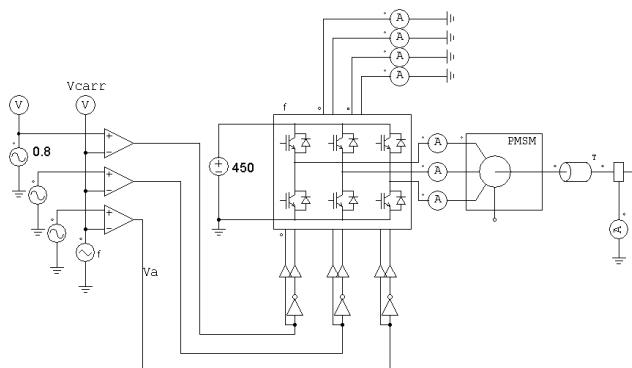


Fig. 5. Layout of test circuit in PSIM

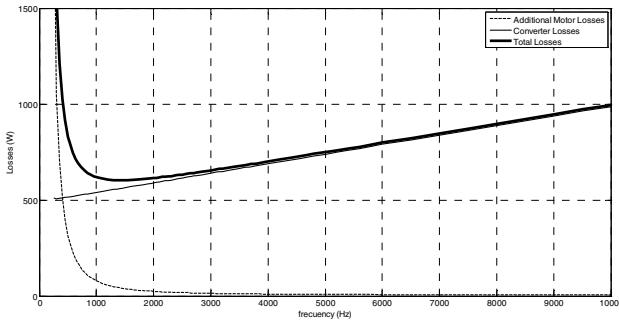


Fig. 6. Total loss of inverter circuit / motor

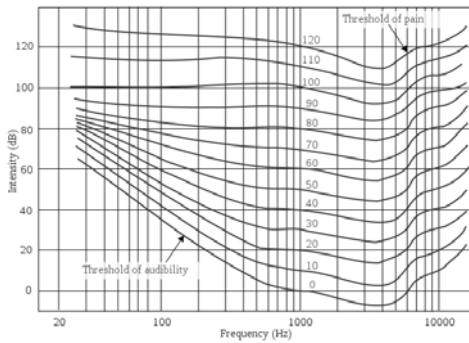


Fig. 7. Human Ear Frequency Range (Fletcher–Munson curves)

In the Fig. 7 the lower curve gives the faintest sounds that can be heard, and the upper curve gives the loudest sounds that can be heard without pain. Taken into account the results of the Fig. 6, a switching frequency of 1 kHz assures that the application does not produce a high impact on human users due to reduction of engine noise of 10 dB [12], (especially in vehicle applications). The commercial servo amplifier used for the modeled motor recommends the operation with a switching frequency of 8 kHz [13], which will produce, (in accordance with Fig. 6), a total losses increment of nearly 32% of the minimum (reached with 1.35 kHz).

7 Conclusions

In this paper it has been discussed some parameters that affect the efficiency of a PMSM, that could be used in electric vehicle applications and will affect its autonomy. It has been shown how the PI coefficient used in the speed controller affect the energy needed to do the transient states (acceleration and braking in vehicle application), being necessary an input energy up to 20% if these coefficients are not selected correctly.

Also, it has been evaluated the total influence of the switching frequency, taken into consideration not only the converter losses, but also the additional losses that the switching components cause in the motor. For the commercial motor modeled in this paper, the switching frequency that minimized these total losses is near 1 kHz that is a value much lower than the value usually recommended by manufacturers. By selecting this low switching frequency a total losses reduction of nearly 30% can be achieved.

Acknowledgments. This work has been developed under the project City-Elec supported by the Ministry of Science and Innovation from the Government of Spain.

References

1. Hill, R.J.: DC and AC Traction Motors. In: IEEE Conference, 9th Institution of Engineering and Technology Professional Development Course on Electric Traction Systems, pp. 33–52 (2007)
2. Chan, C.C., Chau, K.T.: An Overview of Power Electronics in Electric Vehicles. *IEEE Transactions on Industrial Electronics* 44, 3–13 (1997)
3. Pillay, P., Krishnan, R.: Control characteristics and Speed Controller Design for a High Performance Permanent Magnet Synchronous Motor Drive. *IEEE Transactions on Power Electronics* 5, 151–159 (1990)
4. Chunyuan, B., Shuangyan, R., Liangyu, M.: Study on Direct Torque Control of Super High-speed PMSM. In: IEEE International Conference on Automation and Logistics, pp. 2711–2715 (2007)
5. Szabo, C., Incze, I.I., Imecs, M.: Voltage-Hertz Control of the Synchronous Machine with Variable Excitation. In: IEEE International Conference on Automation, Quality and Testing, Robotics, vol. 1, pp. 298–303 (2006)
6. Foo, G.H.B., Rahman, M.F.: Direct Torque Control of an IPM-Synchronous Motor Drive at Very Low Speed Using a Sliding-Mode Stator Flux Observer. *IEEE Transactions on Power Electronics* 25, 933–942 (2010)
7. Kennel, R.: Encoderless Control of Synchronous Machines with Permanent Magnets - Impact of Magnetic Design. In: IEEE International Conference on Optimization of Electrical and Electronic Equipment, pp. 19–24 (2010)
8. Stirban, A., Boldea, I., Andreeșcu, G.-D., Iles, D., Blaabjerg, F.: Motion Sensorless Control of BLDC PM Motor with Offline FEM Info Assisted State Observer. In: IEEE International Conference on Optimization of Electrical and Electronic Equipment, pp. 321–328 (2010)
9. Iepure, L.I., Boldea, I., Andreeșcu, G.D., Iles, D., Blaabjerg, F.: Novel Motion Sensorless Control of Single Phase Brushless D.C. PM Motor Drive, with Experiments. In: IEEE Int. Conf. Optimization of Electrical & Electronic Equipment, pp. 329–336 (2010)
10. Paicu, M.C., Boldea, I., Andreeșcu, G.D., Blaabjerg, F.: Very Low Speed Performance of Active Flux Based Sensorless Control: Interior Permanent Magnet Synchronous Motor Vector Control Versus Direct Torque and Flux Control. *IET Electric Power Applications* 3, 551–561 (2009)
11. Boldea, I., Paicu, M.C., Andreeșcu, G.-D.: Active Flux Concept for Motion Sensorless Unified AC Drives. *IEEE Transactions on Power Electronics* 23, 2612–2618 (2008)
12. Verheijen, E., Jabben, J.: Effect of Electric Cars on Traffic Noise and Safety. National Institute for Public Health and Environment of Netherlands (2010)
13. Mavilor Express Magazine (September 2003)

A DC-DC Step-Up μ -Power Converter for Energy Harvesting Applications, Using Maximum Power Point Tracking, Based on Fractional Open Circuit Voltage

Carlos Carvalho¹, Guilherme Lavareda², and Nuno Paulino³

¹ Instituto Superior de Engenharia de Lisboa (ISEL – ADEETC), Instituto Politécnico de Lisboa (IPL), Rua Conselheiro Emídio Navarro, nº1, 1949-014 Lisboa, Portugal
cfc@isel.ipl.pt

² Departamento de Ciências dos Materiais e UNINOVA/CTS, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Campus FCT/UNL, 2829-516 Caparica, Portugal
gal@fct.unl.pt

³ UNINOVA/CTS, Departamento de Engenharia Electrotécnica, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Campus FCT/UNL, 2829-516 Caparica, Portugal
nunop@uninova.pt

Abstract. A DC-DC step-up micro power converter for solar energy harvesting applications is presented. The circuit is based on a switched-capacitor voltage tripler architecture with MOSFET capacitors, which results in an area approximately eight times smaller than using MiM capacitors for the $0.13\mu\text{m}$ CMOS technology. In order to compensate for the loss of efficiency, due to the larger parasitic capacitances, a charge reutilization scheme is employed. The circuit is self-coded, using a phase controller designed specifically to work with an amorphous silicon solar cell, in order to obtain the maximum available power from the cell. This will be done by tracking its maximum power point (MPPT) using the fractional open circuit voltage method. Electrical simulations of the circuit, together with an equivalent electrical model of an amorphous silicon solar cell, show that the circuit can deliver a power of $1132\ \mu\text{W}$ to the load, corresponding to a maximum efficiency of 66.81%.

Keywords: Electronics, CMOS circuits, Energy Harvesting, Power management circuits, Maximum Power Point Tracking, Amorphous Silicon Solar Cell.

1 Introduction

There is an emerging need to power applications in an autonomous fashion. This need results from the fact that it may not be practical to plug the device to the power grid, nor to use batteries, as they need to be replaced when their charge is depleted. A solution to this problem, that is being increasingly used, is to power the application by collecting the energy that exists in the surrounding environment; this is known as energy harvesting, or energy scavenging, and has been growing in importance. By employing energy harvesting, circuits can virtually operate permanently. As such, there is no need to plug the circuit to the power grid, or to power it by using batteries.

Energy can be harvested from different sources: light (solar [1], [2] or artificial [3]), mechanical [2] (usually vibrations), thermal gradients [2], or electromagnetic [1]. Each of these energy sources has its own advantages and drawbacks, but they all share a common limitation, which is low energy density. This means that the available power for a small energy harvesting powered system will be limited. Amongst of all, ambient light has the highest energy density when compared to other possible ambient energy sources [1]. The power and the voltage produced by a solar cell vary with the connected load and with the amount of incident light. Thus, it is necessary to increase the voltage supplied by a single solar cell to an acceptable value by most circuits (at least, 1 V). So, it is necessary to use a step-up power converter circuit, which in this paper, is based on a switched-capacitor voltage tripler architecture, using $0.13\mu\text{m}$ CMOS technology MOSFET capacitors. This circuit uses an asynchronous state machine to produce a variable frequency clock, regulating the input voltage of the converter (output of the solar cell) to a nearly constant value, corresponding to the maximum power point (MPP). The objective is to dynamically adjust the working input voltage to the MPP voltage value. The output voltage value is maximized when this occurs. This step-up converter circuit tries to harvest as much energy as possible out of the solar cell using a maximum power point tracking (MPPT) approach. There are many MPPT approaches, varying with the availability of resources and the intended application [4]. Amongst these approaches, the Fractional Open Circuit Voltage (Fractional V_{oc}), is the one used in this work.

The research question associated to this work, is if it is possible to join an a-silicon solar cell to a voltage tripler and an MPPT method, and to get a reasonable efficiency performance. The hypothesis for such a combination is explored in the present paper.

2 Contribution to Sustainability

As the purpose of this system is based on energy harvesting, it can operationally contribute for environmental sustainability. The energy used to power the circuit and the energy that the converter provides, besides being non-polluting, it is also free. Moreover, by excluding the use of batteries, there is no need to dispose them of, avoiding additional pollution, nor the use of additional chemicals than those used to manufacture the circuit itself.

3 Novel Results, Contributing to Technological Innovation

The work described in this paper is a fundamental block of a self-powered system using energy harvesting, to be implemented in $0.13\mu\text{m}$ CMOS technology. There are some innovative aspects in the present work, which include the use of cheaper solar cells (made from amorphous silicon), the combination of NMOS and PMOS devices to implement a voltage step-up regulator and the use of a local supply module to power a phase generator. This local supply strategy allows for a more robust command over the switches in the main step-up converter section.

The ability of being self-powered is very important for electronic systems that are intended to monitor and gather information, in locations where it is difficult or even impractical, to obtain energy by normal methods. In such inaccessible places, an energy independent system, with low installation and operation costs, enhances the

benefits of the energy harvesting facet, widening or even opening new possibilities of applications.

4 Electrical Model of the Amorphous Silicon Solar Cell

An a-Silicon photovoltaic cell was built by depositing amorphous silicon with a structure p/i/n on a glass previously covered with ITO (Indium Tin Oxide). The ITO was deposited using rf-PERTE (radio-frequency Plasma Enhanced Reactive Thermal Evaporation) and had a sheet resistance of $35 \Omega/\square$. The active p-type, intrinsic and n-type layers were deposited using PECVD (Plasma Enhanced Chemical Vapor Deposition) and had a thickness of 150\AA , 4500\AA and 500\AA respectively. The frontal aluminum electrode was deposited using thermal evaporation [5]. The solar cell was experimentally characterized and an equivalent electrical model, shown in Fig. 1-a), was obtained.

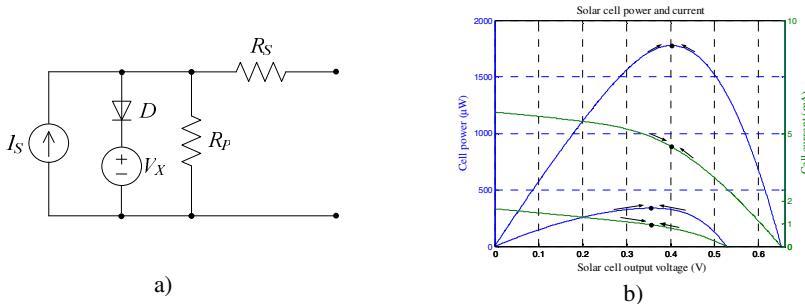


Fig. 1. a) Equivalent electrical circuit of the amorphous silicon solar cell and b) Power and current curves of the solar cell equivalent circuit model for maximum illumination and 30% of maximum illumination (higher and lower curves, respectively)

This solar cell has a short circuit current of about 5.9 mA, a maximum power of $1775 \mu\text{W}$ that occurs for a voltage of 403 mV (maximum power point) and an open circuit voltage of 652 mV. These data refer to a cell having an active area of about 1 cm^2 , when irradiated according to AM1 (Air Mass 1) conditions (irradiance by the solar spectrum at the earth surface, having the Sun vertically located). The resulting power and current curves of the cell are depicted in Fig. 1-b).

Since the MPP voltage is small (around 400 mV) a SC voltage tripler circuit must be used in order to obtain an output voltage value around 1.1V. The impedance that this circuit presents to the solar cell must be adjusted in order for the solar cell voltage to become approximately equal to the MPP voltage.

5 Maximum Power Point Tracking, Based on Fractional Open Circuit Voltage

There are several methods available to track the maximum power point (MPP) of a solar cell [4], in order to achieve efficiencies as high as possible. Some of these

methods can track the true MPP of the cell; however they typically require complex circuits or computational effort, namely the ability to perform multiplications. If some inaccuracy in the determination of the MPP is accepted, it is possible to use simpler methods that require less complex circuits. In this application where the total available power is very low, these simpler methods can be preferable. As such, the Fractional Open Circuit Voltage (Fractional V_{OC}) method was chosen, because it is a very simple and inexpensive (hardware wise) method. This method explores an intrinsic characteristic of cells: there is a proportionality factor between their open circuit voltage and the voltage at which the MPP occurs. This factor must be determined beforehand, by studying the solar cell behavior under several conditions of illumination and temperature. By performing a linear regression over the points plotted on the obtained graphs, the same way as in [6], one can determine the slope of these functions. By sweeping a range of temperatures that spanned from -55°C to $+125^{\circ}\text{C}$, the ratio V_{MPP}/V_{OC} was around 0.84. By sweeping illumination, the ratio V_{MPP}/V_{OC} was around 0.76. Assuming that illumination has more importance, as it is more likely to vary, a value of 0.77 was selected for k , the fractional V_{OC} coefficient. This value agrees with the ones stated in [4].

A pilot solar cell that is in open circuit (unloaded), is used to measure the open circuit voltage. The optimum voltage of the loaded solar cell (MPP), is determined by multiplying the open circuit pilot voltage by k , using a resistive divider. This voltage is known as the fractional V_{OC} . Resistors must be high enough to preserve the open circuit assumption for the pilot cell. The pilot solar cell can be smaller than the main solar cell and it must have the same temperature and illumination as the main cell, in order for the fractional V_{OC} to accurately track the MPP voltage. The MPP tracking is implemented by adjusting the switching frequency of the SC voltage tripler. When the fractional V_{OC} is larger than the solar cell voltage this means that the impedance of the SC circuit is small and therefore it is necessary to decrease the switching frequency in order to increase the impedance, thus increasing the solar cell voltage. If the V_{OC} voltage is smaller than the solar cell voltage, it is necessary to increase the switching frequency in order to decrease the impedance of the SC circuit and therefore decrease the solar cell voltage. This process will result in a switching frequency value that allows the SC voltage tripler circuit to have an impedance value that causes the MPP voltage in the solar cell, as illustrated by the arrows in the graph depicted in Fig. 1-b).

6 Switched Capacitor Voltage Tripler Circuit with Charge Reusing

The circuit of the SC DC-DC converter is shown in Fig. 2. The principle of operation of this circuit is the same of the switched-capacitor voltage tripler [7]. During phase ϕ_1 , the MOS capacitors of the upper half circuit, M_1 and M_3 , are charged with the input voltage value (v_{in}) and then, during phase ϕ_3 , they are connect in series with the input voltage source.

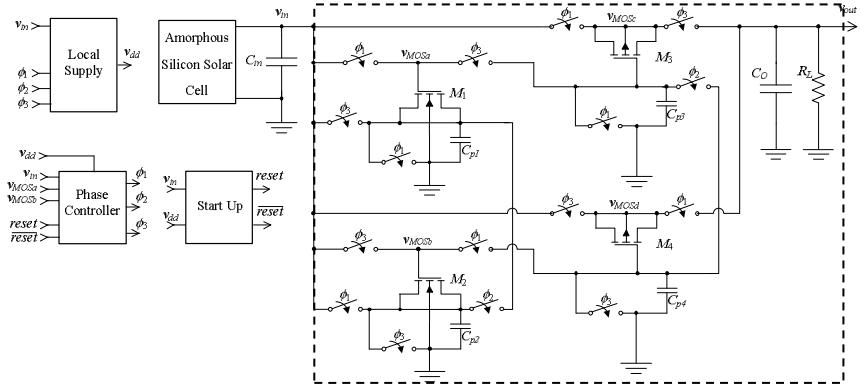


Fig. 2. Schematic of the SC voltage tripler circuit with charge re-utilization

If there were no losses, this would result in an output voltage (v_{out}) approximately three times larger than the input voltage value. Although using MOS capacitors instead of MiM capacitors results in a significant reduction of the occupied area, there is also an increase in the parasitic capacitances of the bottom plate nodes, leading to a decrease in the efficiency of the circuit. In order to reduce the amount of charge lost in these parasitic capacitances, the circuit is split into two halves. The top half is composed by M_1 and M_3 and the bottom half is composed by M_2 and M_4 . The bottom half works in the same way as the top half, with phase ϕ_1 changed with phase ϕ_3 . During an intermediate phase (ϕ_2), the bottom plate nodes of both MOS capacitors of the upper and lower half-circuits are connected together, thus transferring half of the charge in one parasitic capacitance to the other, before the bottom nodes are shorted to ground. This reduces by half the amount of charge that is lost in the parasitic capacitance nodes. The clock phases are generated by the phase controller circuit that will be described next. The output voltage of the circuit depends on the value of the load resistance (R_L). This means that this voltage cannot be used to power the phase generator. Therefore, a smaller SC voltage tripler, controlled by the same clock, is used to create a local power supply. This circuit is a replica of the one inside the dashed rectangle in Fig. 2, but scaled to 3% of its area, as this ratio yielded the best results.

7 Phase Generation and Control

The three clock phases necessary for the operation of the SC voltage tripler circuit are generated by an ASM circuit that automatically adjusts the clock frequency in order to obtain the MPP voltage from the solar cell. This circuit is depicted next in Fig. 3. The operation of this circuit is similar to the one described in [8]. This circuit has four states that are determined by the output of four latches. These states correspond to the clock phases ϕ_1 , ϕ_2 , ϕ_3 , and again ϕ_2 . In order to change from one state to the next, the *Set* signal of one latch is activated, changing the output of that latch from 0 to 1. This, in turn, activates the *Reset* signal of the previous latch, causing its output to change from 1 to 0, thus completing the state change.

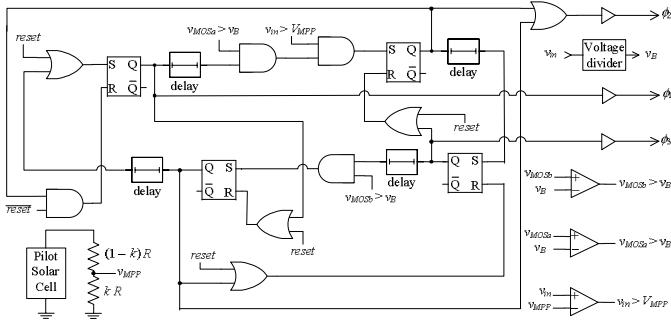


Fig. 3. Phase controller schematic

A start-up circuit (described in [8]) generates a *reset* signal that guarantees that the first state is state1. The ASM is continually changing from one state to the next (and then from state4 to state1), in order to create the clock phases. The maximum frequency of operation is defined by the delay circuits (described in [8]) inserted between the output of each latch and the *Set* input of the next latch. The transitions between state, state2, state3 and state4 are delayed by comparators that guarantee that the MOS capacitors connected to the solar cell are charged to at least 95% of the input voltage ($V_{MOSa} > V_B$ and $V_{MOSb} > V_B$). The duration of state1 (phase ϕ_1) is also dependent on the comparison between the solar cell voltage (v_{in}) and the fractional V_{OC} , obtained from the pilot solar cell (v_{MPP}). When the capacitors are connected to the solar cell (in the beginning of ϕ_1), the voltage v_{in} drops a little. The time it takes to recover to its previous value, and the fractional V_{OC} , depends on the temperature and illumination. Therefore the duration of the period corresponds to the frequency value that adjusts the input voltage to the MPP voltage value.

8 Simulation Results

This step-up converter was designed to work with an amorphous solar cell, with an area of about 1 cm^2 , able to supply a maximum power of $1775 \mu\text{W}$, at a MPP voltage of 403 mV . The efficiency and power values of the circuit for different load resistance values obtained through electrical simulations in Spectre, are shown in Fig. 4-a). This graph shows that when the load resistance is $1.05\text{k}\Omega$, the maximum efficiency of the circuit is 66.81%, for a power delivered to the load of $1132 \mu\text{W}$ and a solar cell power of $1694 \mu\text{W}$. In this situation, v_{in} converged to 450 mV . This value does not match the optimal input voltage of 403 mV , because the Fractional V_{OC} method may not reach the true MPP voltage, as it depends upon k , which is an average value obtained from the studied situations taken beforehand. This problem is not very important since the power available from the cell does not change significantly around the MPP voltage. From a theoretical point of view, as stated in [7], the maximum achievable efficiency of an ideal converter with an input voltage of 450 mV and an output voltage of 1.090 V (which was the value of v_{out} in this situation) would be 80.74%.

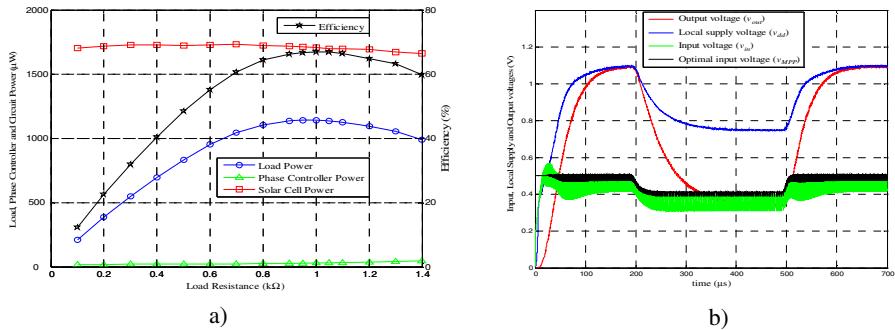


Fig. 4. a) Efficiency, input power, output power and phase controller circuit power as a function of the power delivered to load and b) Evolution of v_{in} , v_{dd} , v_{out} and v_{MPP} , during start-up and transient operation (lower illumination between 200 and 500μs)

In this situation, the phase generator circuit dissipates 32.47 μ W. The frequency of the clock phase ϕ_1 in this load condition is 1.245 MHz. In order to determine how much is the efficiency penalty by using MOS transistors, instead of MiM capacitors, a simulation where the MOS transistors were replaced by MiM capacitors of the given technology was performed. In this case the maximum efficiency increased to 74.59%, which is not significant, given the die area tradeoff. The evolution of the input, local supply, output and optimal input voltages, during start-up (and transient operation), is shown in Fig. 4-b).

Electrical simulations showed that the system can converge, in order for the cell to provide the voltage at which the MPP is achieved. This system can start-up with loads starting as low as 100 Ω . It is possible to have such a significant load connected to the output during start-up because the load of the local supply module is not very significant, allowing for the phase controlling signals to be correctly defined.

In order to check the robustness of the MPP tracker circuit when experiencing a sudden illumination change, Fig. 4-b) also shows how the circuit behaves under a irradiance-transient operation, when illuminated by 100% and 30% of the maximum value. It is seen that the circuit is able to track the optimal input voltage in both situations. Clock signal ϕ_1 decreases its frequency when a lower irradiance level is present, because the amount of available charge from the solar cell is also lower.

9 Conclusions

A step-up micro-power converter for solar energy harvesting applications was presented. The circuit is based on a switched-circuit voltage tripler architecture, with MOSFET capacitors of the 0.13 μ m CMOS technology, which results in a total circuit area approximately eight times smaller than using MiM capacitors of the given technology. In order to compensate for the loss of efficiency due to the larger parasitic capacitances, a charge reutilization scheme was employed. The circuit uses a phase controller, designed specifically to work with an amorphous solar cell, in order to track the MPP of the cell, using the fractional V_{OC} method. To implement this method, a

previous study of the cell characteristics must be carried out, regarding illumination and temperature. The controller is powered by a local supply circuit to ensure that the phase signals that control the main switches are well defined. Electrical simulations of the circuit together with an equivalent electrical model of the amorphous solar cell, have shown that the circuit can deliver a power of 1132 μ W to the load and a total circuit power dissipation of 1694 μ W, corresponding to a maximum efficiency of 66.81%. This efficiency value is similar to the one obtained in [9], meaning that the hypothesis proposed in this paper is able to meet the requirements formulated in the research question. When using the MiM capacitors of the 0.13 μ m CMOS technology, the increase of efficiency to 74.59% is not significant, considering the eight times less die area tradeoff. When the solar cell experiences irradiance variations, the phase controller circuit can effectively track the MPP, as it shifts from its previous value.

Acknowledgments. This work was supported by the Portuguese Foundation for Science and Technology (FCT/MCTES) (CTS multiannual funding) through the PIDDAC Program funds.

References

- Chalasani, S., Conrad, J.M.: A survey of energy harvesting sources for embedded systems. In: Southeastcon, April 3-6, pp. 442–447. IEEE, Los Alamitos (2008)
- Paradiso, J.A., Starner, T.: Energy scavenging for mobile and wireless electronics. In: Pervasive Computing, vol. 4(1), pp. 18–27. IEEE, Los Alamitos (2005)
- Hande, A., Polk, T., Walker, W., Bhatia, D.: Indoor solar energy harvesting for sensor network router nodes. Microprocessors and Microsystems 31(6), 420–432 (2007)
- Esram, T., Chapman, P.L.: Comparison of Photovoltaic Array Maximum Power Point Tracking Techniques. IEEE Transactions on Energy Conversion 22(2), 439–449 (2007)
- Amaral, A., Lavareda, G., Nunes de Carvalho, C., Brogueira, P., Gordo, P.M., Subrahmanyam, V.S., Lopes Gil, C., Duarte Naia, V., de Lima, A.P.: Influence of the a-Si:H structural defects studied by positron annihilation on the solar cells characteristics. Thin Solid Films 403-404, 539–542 (2002)
- Brunelli, D., Moser, C., Thiele, L., Benini, L.: Design of a Solar-Harvesting Circuit for Batteryless Embedded Systems. IEEE Transactions on Circuits and Systems – I: Regular Papers 56(11), 2519–2528 (2009)
- Zhu, G., Ioinovici, A.: Switched-capacitor power supplies: DC voltage ratio, efficiency, ripple, regulation. In: Proc. IEEE International Symposium on Circuits and Systems ISCAS 1996. Connecting the World, pp. 553–556 (1996)
- Carvalho, C., Paulino, N.: A MOSFET only, step-up DC-DC micro power converter, for solar energy harvesting applications. In: Proceedings of the 17th International Conference Mixed Design of Integrated Circuits and Systems (MIXDES), June 24-26, pp. 499–504 (2010)
- Shao, H., Tsui, C.-Y., Ki, W.-H.: The Design of a Micro Power Management System for Applications Using Photovoltaic Cells With the Maximum Output Power Control. IEEE Transactions on Very Large Scale Integration (VLSI) Systems 17(8), 1138–1142 (2009)

Wireless Sensor Network System for Measuring the Magnetic Noise of Inverter-Fed Three-Phase Induction Motors with Squirrel-Cage Rotor

Andrei Negoita¹, Gheorghe Scutaru¹, Ioan Peter², and Razvan Mihai Ionescu¹

¹ Transilvania University of Brasov, Advanced Electrical Systems Department,
Politehnicii Street No.1, 500036 Brasov, Romania

{Andrei.Negoita,Gheorghe.Scutaru,Razvan.Mihai.Ionescu}

andrei.negoita@yahoo.com

² S.C. Electroprecizia S.A., Motor Design Department,
Parcului Street No. 18, 505600 Brasov, Romania

{Ioan.Peter}pr_mot@electroprecizia.ro

Abstract. The object of this paper is the study of the noise produced by inverter-fed three-phase induction motors with squirrel-cage rotor. A wireless sensor network based measurement system is proposed, which gives the possibility of measuring the sound pressure virtually simultaneously in multiple points around the motor. In the case of inverter fed motors, the phenomena that lead to the production of the magnetic noise become more complex and the motor becomes noisier because of the increased possibility of matching the exciting frequencies with stator natural frequencies. In order to evaluate the influence of the switching frequency of the PWM inverter on the overall motor noise, the noise-frequency level diagrams (spectrograms) have been traced for a two speed motor of 1.5/2.4 kW, 750/1500 rpm, with 36 stator slots and 46 rotor slots.

Keywords: wireless sensor network, induction motor, squirrel-cage, noise.

1 Introduction

One of the main sources of noise in a rotating electrical machine is the vibrations excited by electromagnetic forces acting in the motor air-gap. The use of inverters for controlling the speed of induction motors leads to an increase in the harmonic content of the supply voltage. As a consequence, the harmonic content of the air-gap flux increases, thus creating a larger number of magnetic force waves. If the natural frequencies of the motor structure match the frequencies of the magnetic forces, the resonance phenomenon appears which contributes greatly to an increase of the overall noise level of the motor [1, 2].

2 Contribution to Sustainability

In our days, inverter fed induction motors are frequently used in the residential field. As contribution to the sustainable development, design techniques are required to assure a reduced noise pollution level.

The object of this paper is the study of the noise produced by inverter-fed three-phase induction motors with squirrel-cage rotor. A wireless sensor network system is proposed for measuring the sound pressure virtually simultaneously in multiple points around the motor. The measuring system can be implemented by using low-cost, low power sensors such as miniaturized condenser microphones.

A solution for implementing a system containing a large number of sensors requires the development of a wireless sensor network (WSN) [3]. The flexibility, fault tolerance, high sensing fidelity, low-cost and rapid deployment characteristics of such WSN makes them an ideal platform for condition monitoring of electrical machines[4]. WSN have been applied to condition monitoring of induction motors, either by using the motor current spectral analysis technique which requires the stator current to be sampled and collected [5] or by monitoring bearing vibrations by wireless accelerometers [6].

The IEEE 802.15.4 communication protocol, allows small, power efficient and inexpensive solutions to be implemented for a wide range of devices [7], such as wireless sensors. The protocol can be used for implementing WSN architectures for low cost applications with data throughput as a second consideration. The IEEE 802.15.4 protocol is intended to address applications wherein existing wireless solutions are too expensive or difficult to implement. Table 1 compares the performance of different wireless technologies.

Table 1. A comparison of the 802.15.4 standard with other wireless technologies

	802.11b WLAN	802.15.1 Bluetooth	802.15.4 ZigBee
Range [m]	100	10 - 100	10
Data Throughput [Mbps]	11	1	0.25
Complexity	High	High	Low
Cost	High	Medium	Low

The 802.15.4 standard allows the formation of two possible network topologies: the star topology and the peer-to-peer topology as seen in Figure 1.

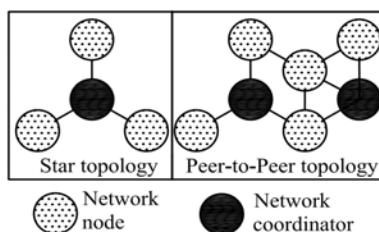


Fig. 1. IEEE 802.15.4 network topologies

One of the main problems concerning the implementation of a WSN network consists in finding the best possible way in which to combine the diverse and sometimes conflicting data gathered by the nodes. Multi-sensor information fusion

can increase measurement credibility thus improving system reliability. There are many methods employed in multi-sensor data fusion such as [8]: Kalman filter, Bayes estimation, fuzzy set theory and neural networks. The implementation of a Kalman filter approach in the case of the proposed measurement system will be discussed in a future paper.

For the proposed system, a ZigBee protocol WSN was implemented using the Microchip Stack for the ZigBee Protocol. The protocol uses the IEEE 802.15.4 specification as its Medium Access Layer (MAC) and Physical Layer (PHY). According to the IEEE 802.15.4 standard, three types of devices exist in a network.

Their main functions are summarized in Table 2.

Table 2. ZigBee and IEEE 802.15.4 standard device types

ZigBee device type	IEEE device type	Network function
Coordinator	Full Function Device	Forms the network, allocates network addresses or allows other devices to join the network
Router	Full Function Device	Optional device which extends the physical range of the network or performs monitoring and control functions.
End	Full Function or Reduced Function Device	Performs monitoring and control functions.

As seen in Figure 1, depending on the chosen topology, the elements of a WSN network can communicate with each other directly or through the coordinator. There are two possible types of multi-access mechanisms: beacon and non-beacon. In a non-beacon enabled network, all nodes in the network are allowed to transmit at any time if the channel is idle. In a beacon enabled network, nodes are allowed to transmit in predefined time slots only. On power-up, the protocol coordinator, based on the number of networks found on each allowed channel, establishes its own network and selects a unique 16-bit Personal Area Network identification. Once a new network is established, protocol routers and end devices are allowed to join the network.

For the proposed system a star topology, 2.4 GHz, 250kbps, non-beacon, network using a single protocol coordinator was implemented using the Microchip Stack for the ZigBee Protocol.

A sound measurement WSN sensor node consists of a signal conditioning circuit, a microcontroller for data acquisition and conversion, a memory module for data storage and a wireless communication module. Thus the node has both sensing and communication capabilities. The simplified schematic is shown in Figure 2.

The signal conditioning block filters and amplifies the signal coming from the condenser microphone. An inverter amplifier is implemented, using R4 and R5 to set the amplifier offset voltage and R3 and R2 to control the gain. R6 and C4 form an anti-alias low pass filter with a cut-off frequency of 38 kHz.

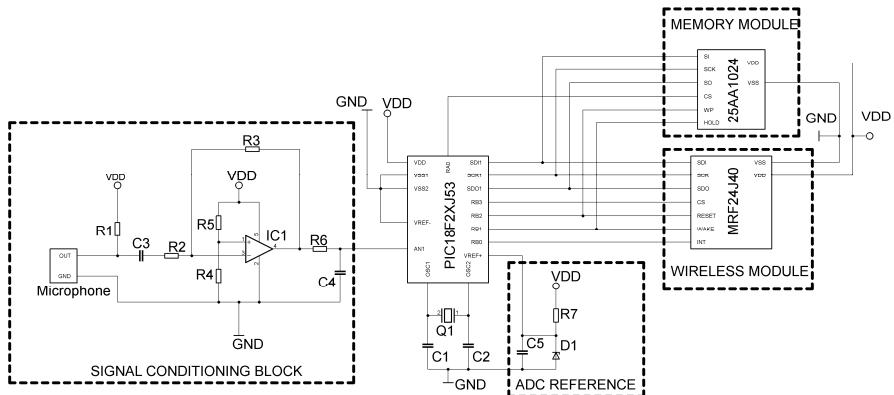


Fig. 2. WSN sensor node simplified schematic

One of the main concerns when signal processing is involved is the aliasing effect, i.e. frequency components of the acquired analog signal greater than half the sample rate of the analog-to-digital converter that shift into the frequency band of the output signal, thus distorting it. Therefore, the sampling frequency of the PIC18F27J53 microcontroller analog-to-digital module must be at least twice the highest frequency of interest of the sampled signal. In our case the sampling frequency was set at 77 kHz.

As shown in Figure 3, the microcontroller uses a 12 bit analog-to-digital module for converting the analog signal. The data is then stored into a 1Mbit serial EEPROM memory. The data acquisition process continues as long as the duration of the measurement process set by the user has not been achieved. When the measurement process has ended, the microcontroller begins reading the stored data and sending it to the network coordinator by using the MRF24J40 wireless module.

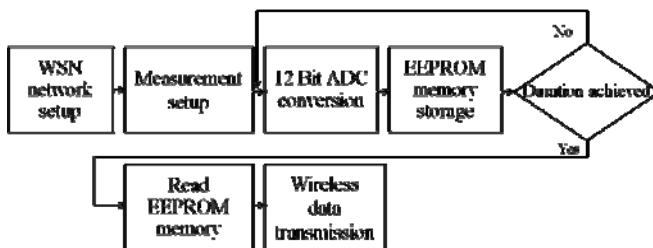


Fig. 3. Data processing and transmission flow-chart

A WSN receiver node, in our case, the network coordinator, receives data sequentially from each of the sensor nodes on the network. As each node transmits, the data is converted by the interface microcontroller and sent to a PC for processing.

The sound measurements were conducted in a semi-anechoic room in compliance with the ISO 1680/1 standard. The WSN nodes can be placed around the motor, on the measurement positions defined by the ISO 1680/1 standard. The measurement setup is shown schematically in Figure 4.

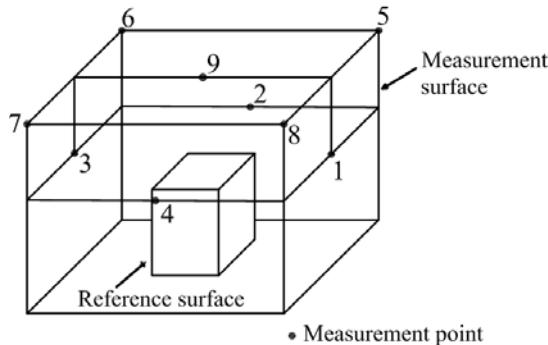


Fig. 4. Measurement points around the motor according to ISO 1680/1 standard

3 The Magnetic Noise of Inverter-Fed Induction Motors

Inverter-fed induction motors are noisier than those fed with sinusoidal current because of the increased possibility of matching the magnetic forces exciting frequencies with stator natural frequencies. The order of the stator current harmonics of an inverter-fed three phase induction motor is:

$$n = 6k \pm 1 \quad (1)$$

By neglecting the tangential component of the magnetic flux density, according to the Maxwell stress tensor, the magnetic pressure waveform at any point of the air gap can be expressed as:

$$p_r(\alpha, t) = \frac{b^2(\alpha, t)}{2\mu_0} = \frac{1}{2\mu_0} [b_1(\alpha, t) + b_2(\alpha, t)]^2 \quad (2)$$

In terms of Fourier series, the following groups of magnetic waves are produced [1]:

- $p_{rvn}(\alpha, t)$, determined by the product $[b_1(\alpha, t)]^2$ of the stator harmonics having the same order v . The frequency of the radial magnetic pressure is $f_{rn} = 2nf$ and the vibration mode $r = 2v p$.
- $p_{r\mu n}(\alpha, t)$, determined by the product $[b_2(\alpha, t)]^2$ of the rotor harmonics having the same order μ . The frequency of the radial magnetic pressure is $f_{rn} = 2nf_\mu$ and the vibration mode $r = 2\mu p$.
- $p_{rv\mu n}(\alpha, t)$, determined by the interaction of stator harmonics having the order v and rotor harmonics having the order μ . The frequency of the radial magnetic pressure is $f_{rn} = n(f \pm f_\mu)$ and the vibration mode $r = (v \pm \mu)p$.

Higher time stator harmonics of different numbers can produce significant radial forces [2] having the frequency $f_{rn} = (n' \pm n'')f$; $n' \neq n''$ and the vibration mode

$r = 0$ or $r = 0$. The most important are the magnetic forces due to sums and differences of the fundamental harmonic f with higher order time harmonics of the stator current [1]:

$$f_{r n} = (1 \pm n)f \quad (3)$$

The inverter switching frequency has an important effect as the interaction of switching frequency harmonics and higher time harmonics, produces forces with frequencies $f_n = n' f_{sw} \pm n'' f$. If n' is an odd integer, n'' will be an even integer and vice versa:

$$f_n = f_{sw} \pm 2f, \dots \quad \text{or} \quad f_n = 2f_{sw} \pm f, \dots \quad (4)$$

Significant vibration can result from the interaction of permeance field harmonics and MMF harmonics associated with higher time harmonics of the stator [2]:

$$f_{r n} = \left| f_n \pm f \left[1 + k \frac{s_2}{p} (1-s) \right] \right| \quad (5)$$

and the vibration mode $r = 0, 2$.

In the case of inverter-fed motors, the magnetic component of noise is modified. In order to evaluate this modification, a two speed, 1.5/2.4 kW, 750/1500 rpm motor, with 36 stator slots and 46 rotor slots was tested. The spectrograms have been traced using a Brüel & Kjaer 2112 spectrum analyzer.

The motor was tested for no load operation and was supplied from the network or a "Telemecanique" ATV-58HU54N4 inverter with a switching frequency set at 2 kHz. The results obtained for the 1500 rpm speed are shown in Figure 5 and 6.

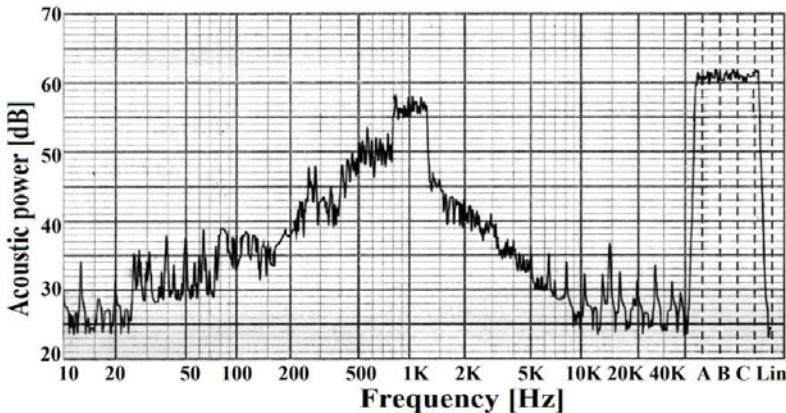


Fig. 5. No load operation, $U_N=400V$ and 50 Hz, network supplied, 2.4 kW-1500 rpm

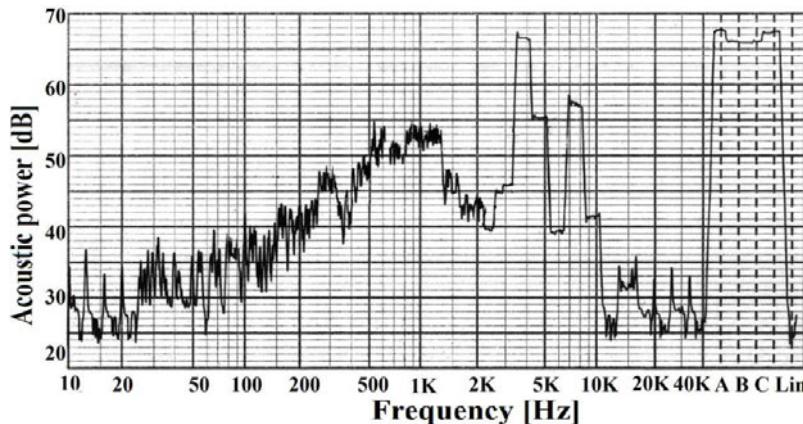


Fig. 6. No load operation, $U_N=400\text{V}$ and 50 Hz, inverter supplied, 2.4 kW-1500 rpm

For the 1500 rpm speed, two new noise peaks appear between 2500 and 4000 Hz. and 7000 and 8000 Hz. The noise increases by 7 dB.

4 Conclusions and Future Work

The presented spectrograms will be used as a reference for evaluating the performance of the proposed wireless sound measuring system.

The applicability of the system depends on several factors. Battery life is extremely important, as most power is used for wireless communication. In order to maximize battery life, the system must take advantage of the sleep mode operation feature of the microcontroller and the wireless communication modules.

The overall speed and accuracy of the proposed system can be significantly improved by using larger data throughput protocols like Bluetooth and superior microprocessors like Digital Signal Processors. From the implementation point of view, this could eliminate the need for an external EEPROM memory and would provide real time sound measurements capability. However, these improvements would lead to significant increase of system cost and development time.

Acknowledgement

This paper is supported by the Sectoral Operational Programme Human Resources Development (SOP HRD), ID59321 financed from the European Social Fund and by the Romanian Government.

References

1. Scutaru, G., Peter, I.: The noise of electrical induction motors. LUX Libris Publishing House, Brasov (2004) (in Romanian)
2. Gieras, J., Wang, C., Cho Lai, J.: Noise of polyphase induction motors. Taylor & Francis, Abington (2006)

3. Yick, J., Mukherjee, B., Ghosal, D.: Wireless sensor network survey. *Computer Networks* 52, 2292–2330 (2008)
4. Korkua, S., Jain, H., Lee, W., Kwan, C.: Wireless Health Monitoring System for Vibration Detection of Induction Motors. In: IEEE Industrial and Commercial Power Systems Technical Conference, pp. 1–6 (2010)
5. Lu, B., Wu, L., Habetler, T.G., Harley, R.G., Gutierrez, A.T.: On the Application of Wireless Sensor Networks in Condition Monitoring and Energy Usage Evaluation for Electric Machines. In: 31st Annual Conference of the IEEE Industrial Electronics Society, pp. 2674–2679 (2005)
6. Jagannath, V.M.D., Raman, B.: WiBeaM: Wireless Bearing Monitoring System. In: 2nd Int. Conf. on Communication Systems Software and Middleware, pp. 1–8 (2007)
7. IEEE 802.15.4 Standard for Information Technology - Telecommunications and information exchange between systems - Local and metropolitan area networks - Specific requirements, <http://standards.ieee.org/getieee802/802.15.html>
8. Zhou, H.: Multi-sensor Information Fusion Method Based on the Neural Network Algorithm. In: Fifth International Conference on Natural Computation, ICNC 2009, pp. 534–536 (2009)

Axial Disc Motor Experimental Analysis Based in Steinmetz Parameters

David Inácio¹, João Martins¹, Mário Ventim Neves¹, Alfredo Álvarez²,
and Amadeu Leão Rodrigues¹

¹ CTS/UNINOVA, Faculty of Sciences and Technology – University New of Lisbon,
Quinta da Torre, 2829-516 Caparica, Portugal

² Department of Electrical Engineering, Escuela de Ingenierías Industriales,
Universidad de Extremadura, E-06006 Badajoz, Spain

Abstract. Nowadays an economical and environment crisis is felt in the world due to the increasing fuel prices and high CO₂ emissions. This crisis is mostly due to present the transportation system, which uses internal combustion engines. The development and integration of electrical motors with improved electro mechanical characteristics, using high temperature superconductors, can provide a sustainable future replacing the conventional internal combustion motors.

An axial type disc motor, with high temperature superconductor (HTS) material has been developed and tested in order to obtain an electrical equivalent circuit based on the experimental results. The tested HTS motor exhibits a conventional hysteresis motor type of behavior, even though the hysteretic phenomena don't have the same principle. The proposed approach allows the description of the equivalent electrical circuit as a conventional hysteresis motor.

Keywords: Axial disc motor; HTS materials; YBCO.

1 Introduction

Economical, environmental and political issues make the optimization and improvement of electric machines necessary to develop electrical vehicles and ensure a sustainable future. These types of vehicles are developed purely electric (for example, batteries or fuel cell and hydrogen fed) or hybrid (combustion and electric fed), being integrated in various projects.

The HTS materials present some advantages making them unique. Almost null DC resistivity, high current transportation capability and trapping flux capability, enables increasing efficiency several electrical applications [1-4]. These advantages allow the developing of superconducting electrical machines with higher specific torque than their conventional counterparts [5], [11-12].

Electrical machines with HTS bulk rotor present a complex behavior, showing both synchronous and asynchronous regimes [5], [7]. Even though they are similar to conventional hysteresis motors, their operating principle is different. It is appropriate to study the behavior of superconducting hysteresis motors in asynchronous regime using Steinmetz-type models, in order to compare it with the conventional induction motor and to clarify the effect of the HTS bulk material in this regime.

In this paper, the relations between the equivalent circuit's parameters and the motor's characteristics are discussed, for a conventional induction disk motor and for a HTS hysteresis disk motor, composed by a rotor with a polycrystalline YBCO disk [12]. This will be made using the Steinmetz-type "T" model.

2 Contribution to Sustainability

The paper presents an analysis of two types of disc motors: conventional and high temperature superconducting. This analysis, based on experimental parameters and electromechanical characteristics, allows to conclude that the HTS disc motor presents better electromechanical characteristics than the conventional one. The use of HTS disc motors in future full electric vehicles, based on fuel cell power supplies, will contribute to ensure the desired full electric vehicles levels of environmental sustainability.

3 Expression of Equivalent Electrical Circuit

The steady state per phase Steinmetz electrical equivalent circuit of both a HTS hysteresis and an induction motor, considering that the stator's winding number of turns is equal to the rotor's winding number of turns, can be expressed as shown in fig. 1 (see notation in Appendix A2) [8]. Eddy current component was neglected.

In both motors the rotor resistance is separated as a component proportional to the electrical losses, R_A , and as a component proportional to the mechanical output, R_B , shown in fig. 1.

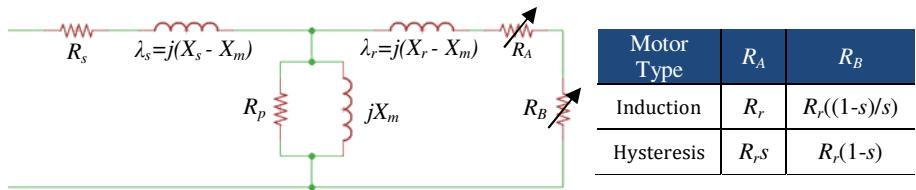


Fig. 1. Equivalent Steinmetz's electrical circuit for a conventional induction and hysteresis motor [8]

To completely specify the Steinmetz equivalent circuit one has to establish its parameters. For the motor equivalent electrical circuit, the stator circuit parameters can be easily determined, directly in the stator. However this could not be done on the rotor side.

$$\begin{aligned}\lambda_s &= (1 - \alpha k) X_s = 0,33 X_s \\ \lambda_r &= (\alpha^2 - \alpha k) X_s = 0,33 X_s \\ X_m &= \alpha k X_s = 0,67 X_s\end{aligned}\tag{1}$$

For the induction motor, these parameters are estimated through the blocked-rotor and no-load tests. For the HTS hysteresis motor, according to [8] and for $\alpha = 1$ (value for

a squirrel-cage type induction motor) and $k = 0.67$ (calculated in [8] for a similar motor), Steinmetz's parameters are given by equation (1).

3.1 Blocked-Rotor, No-Load and Load Tests

For blocked-rotor test the mechanical speed is null, hence $s = 1$, and therefore $R_B = 0$ (for both induction and hysteresis machines), which corresponds to a “virtual” short-circuit in the rotor's side equivalent circuit. At no-load test the mechanical speed is equal to the synchronous speed, hence $s = 0$, leading to a rotor's side open circuit for the induction motor. In each test, the stator winding voltage U , the winding current I , and the electrical power P_{elect} are measured per phase. The per phase complex impedance is given by $\bar{Z} = Z e^{j\varphi}$, where $Z = U/I$ and $\varphi = \cos^{-1}[P_{elect}/(U.I)]$. Longitudinal and transversal complex impedances are, respectively, obtained using the blocked-rotor test ($s = 1$), where the magnetization current is neglecting, and the no-load test ($s = 0$), where the stator impedance is neglecting. Both are presented in (2)

$$\begin{aligned}\bar{Z}_L &= R_{L.eq} + jX_{L.eq} \quad \left\{ \begin{array}{l} R_{L.eq} = |\bar{Z}_L| \cdot \cos(\varphi) = R_s + R_r \\ X_{L.eq} = |\bar{Z}_L| \cdot \sin(\varphi) = X_s + X_r \end{array} \right. ; \quad \bar{Z}_T = R_m + jX_m \quad \left\{ \begin{array}{l} R_m = |\bar{Z}_T| (\cos(-\varphi))^{-1} \\ X_m = |\bar{Z}_T| (\sin(-\varphi))^{-1} \end{array} \right.\end{aligned}\quad (2)$$

Determining the value of R_s by independent measurements, and considering $X_s = X_r$ (which is a usual assumption, based on theoretical considerations and supported on practical evidence [13]), all the parameters of Steinmetz's equivalent electrical circuit can be obtained from the previous relations.

The load test is performed with the aim of studying the motor's behavior. It consists in applying some mechanical load torque, driven by the shaft. From the analysis and computation of torque and speed measurements the motor's mechanical characteristics could be obtained.

4 Power Output and Torque Analysis

Even though HTS materials present a different hysteresis phenomenon, when compared with the ferromagnetics ones, in both of them, AC losses are directly proportional to the hysteresis loop area and to the frequency. The output power and the electromechanical torque, in a HTS machine, are given in table 1 (appendix A1).

Table 1. HTS Hysteresis motor and Induction motor electromechanical characteristics

HTS Hysteresis Machine	$P_{mec} = P_{elect}(1-s) = P_{H/cicle} \cdot P.f_s(1-s)$	$T_{elmech} = \frac{p^2}{2\pi} P_{H/cicle}$
Induction Machine	$P_{mec} = R'_c \cdot I_r^2 = R'_r \left(\frac{1-s}{s} \right) I_r^2$	$T_{elmech} = \frac{p}{\omega_{sup}} \frac{R'_r}{s}$

The conventional induction machine is analyzed, according the most literature, based in the Steinmetz equivalent circuit. The power output and torque equations are disposed in table 1, that permits conclude that, for the HTS motor, the output power is

proportional to the AC losses in the HTS material and to the slip, which means that for values of speed near of synchronous speed, all the AC losses in HTS materials is transformed in mechanical output power.

The analysis of table 2 allows to conclude that, for the HTS motor, the output power is proportional to AC losses in the HTS materials and to the sleep, this means that close to synchronism operation, all HTS materials AC losses are converted into mechanical output power. The electromechanical torque is independent of the slip and directly proportional to HTS AC losses, with a factor of 2π (this result is in line with [5] and [8]). Higher HTS material AC losses imply higher mechanical power and torque.

5 Experimental Results

5.1 Test Method and Experimental Apparatus

In this paper the blocked-rotor, no-load and load tests were performed for a conventional induction motor. The blocked-rotor and load tests were only made for the HTS motor with the main objective of finding the Steinmetz's parameters and observing the behavior of both conventional and superconductor motors.

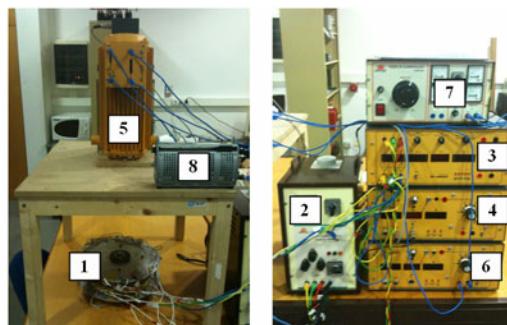


Fig. 2. Experimental apparatus for induction and HTS hysteresis motor's tests

In the performed tests the motors {1} were fed with 3-phase [phase-to-phase rms voltage of 40 V (conventional induction motor) and 23 V (HTS motor)], four poles and a 50 Hz supply frequency configuration. The mechanical load was obtained by means of a DC generator (controlled by two dc power supply and a control resistance) driven by the motors' shaft and feeding a resistive load. The basic used instrumentation is depicted in fig. 2: a transformer {2} to feed the motor; a commercial electrical {3} and mechanical {4} power measuring modules to measures the electrical and mechanical parameters in the system. The DC machine {5} had sensors, which communicate with the modules, to mechanical torque and speed measurement. The DC Generator was controlled with two DC current supply {6}{7} and a control resistance {8}. During the superconductor motor's tests, only the disk motor was immersed in liquid nitrogen. Different values of electrical current is due to limitations in the supply transformer, whose rated current is limited to 20 A.

5.2 Experimental Determination of Steinmetz Equivalent Electrical Circuit

For the presented induction motor, using the equations presented in (2) and from the measured experimental results, in table 2, the parameters were determined. For the HTS hysteresis motor the measured electric quantities were used to compute the longitudinal complex impedance and using (1), the Steinmetz parameters were computed. All the computed values are presented in table 2.

Table 2. Stator's measured and obtained Steinmetz's parameters

	Measured results		Computed results			
	$R_s[\Omega]$	$X_s[\Omega]$	$\lambda_s, \lambda_r [\text{mH}]$	$R_r[\Omega]$	$R_m[\Omega]$	$X_m[\Omega]$
Induction Machine	1,06	0,47	1,50	0,028	2,85	2,2
HTS Hysteresis Machine	0,3	0,27	0,088	0,0008	2,85	0,18

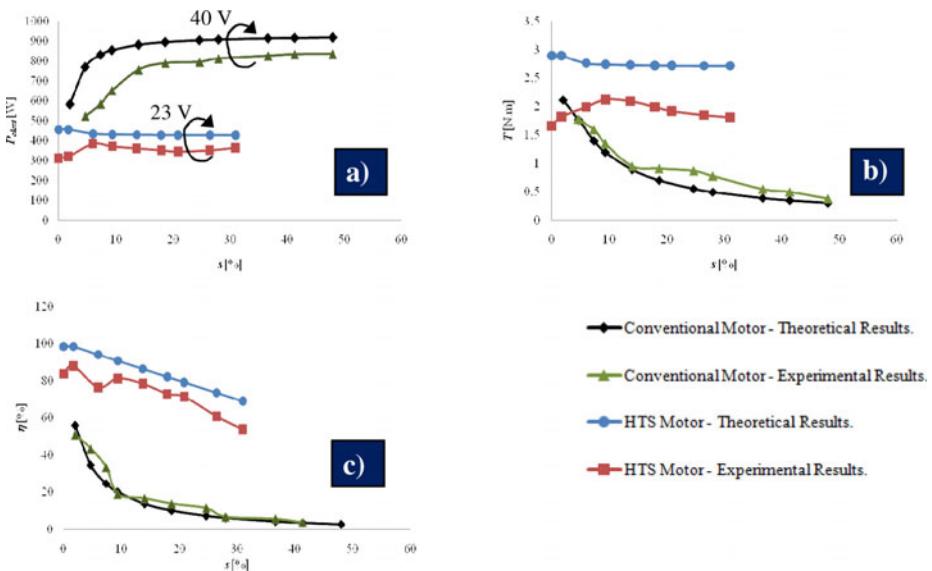


Fig. 3. Comparison between theoretical and experimental a) electrical power, b) developed torque and c) efficiency of conventional induction and HTS hysteresis motor

The value of the magnetic losses, R_m , was calculated based upon the induction motor case and considered the same for the both cases. For the HTS case, doesn't exist information that allows its computation.

The load test was performed as described in section 3. For both motors, the theoretical electromechanical characteristics were computed, based on the experimentally obtained parameters and using the theoretical analysis above described. These were compared with the experimentally obtained characteristics in tests. Various slip-dependent characteristics were present in figure 3: a) one can see that the electrical power is higher than the correspondent value in HTS motor, which

is expected because the supply voltage is different; b) and c) respectively, shows the both motor's torque and the efficiency. The experimental characteristics have a trend evolution coherent with the theoretically predicted ones.

The comparison between predicted and measured quantities must be taken as being essentially indicative, for two reasons. On one hand, the measured quantities, being much smaller than the measuring apparatus' ranges, suffer from an important relative error. And, on the other hand, the predicted quantities refer to the internal power or torque electrodynamically developed, while the measured ones refer to the available quantities in the motor's shaft, which differ from the first ones by the mechanical losses. The viscosity friction in the liquid nitrogen and the friction in the conventional bearings working at low temperature cause non negligible losses when compared with the low quantities developed. Therefore, the measured results for the HTS motor are consistently and significantly lower than the theoretical ones.

6 Conclusions and Future Work

The research & development in the electrical machinery and control areas could provide innovative machines that associated to innovative control methods can be used to achieve a desirable sustainable future. The integration of HTS materials provides better performances in the electrical machines.

Low power prototypes of a conventional induction axial type motor and a HTS axial type motor were tested. The conventional induction motor presents a higher range of measured slip values than the HTS hysteresis motor because with a 30% slip value the current was near to the maximum that the system supply could deliver. Nevertheless the minimum value of slip is high. To obtain more detailed initial values the synchronism speed test must be performed in future work.

The torque ratio between the HTS and the conventional motors isn't as higher as expected. The different supply voltage used can be the reason for the shown behavior. Still, the HTS motor presents a higher efficiency compared with the induction motor. However, further research work must be done to explain the obtained results, with the implementation of an optimized experimental apparatus that includes a supply transformer allowing higher current and including a measurement system with a range adequate for the measured values.

Acknowledgements. Authors would like to thank to CTS of UNINOVA and to FCT (CTS multiannual funding - through the PIDDAC Program funds) for the financial support for this work.

References

1. Kawabata, S., Nakahama, Y., Kawagoe, A., Sumiyoshi, F.: Development of a Compact HTS Current Transformer for Evaluating the Characteristics of HTS Conductors. *IEEE Transactions on Applied Superconductivity* 18(2), 1147–1150 (2008)
2. Noe, M., Juengst, K.-P., Werfel, F., Cowey, L., Wolf, A., Elschner, S.: Investigation of high-Tc bulk material for its use in resistive superconducting fault current limiters. *IEEE Transactions on Applied Superconductivity* 11(1), 1960–1963 (2001)

3. Nomura, S., Tanaka, N., Tsuboi, K., Tsutsui, H., Tsuji-Iio, S., Shimada, R.: Design considerations for SMES systems applied to HVDC links. In: 13th European Conference on Power Electronics and Applications, EPE 2009, pp. 1–10, 8–10 (2009)
4. Ohsaki, H., Tsuboi, Y.: Study on electric motors with bulk superconductors in the rotor. Journal of Materials Processing Technology 108, 148–151 (2001)
5. Barnes, G., Dew-Hughes, D., McCulloch, M.: Finite difference modelling of bulk high temperature superconducting cylindrical hysteresis machines. Supercond. Sci. Technol. 13(2), 229–236 (2000)
6. Inácio, D., Martins, J., Ventim-Neves, M., Álvarez, A., Leão-Rodrigues, A.: Disc motor: Conventional and superconductor simulated results analysis. In: Camarinha-Matos, L.M., Pereira, P., Ribeiro, L. (eds.) DoCEIS 2010. IFIP Advances in Information and Communication Technology, vol. 314, p. 505. Springer, Heidelberg (2010)
7. Inacio, D., Inacio, S., Pina, J., Valtchev, S., Ventim-Neves, M., Martins, J., Leão-Rodrigues, A.: Conventional and HTS disc motor with pole variation control. In: Power Engineering, Energy and Electrical Drives, POWERENG 2009, vol. 18, p. 513 (2009)
8. Jung, H., Nakamura, T., Tanaka, N., Muta, I., Hoshino, T.: Characteristic analysis of hysteresis-type Bi-2223 bulk motor with the use of equivalent circuit. Physica C: Superconductivity 405(2), 117–126 (2004)
9. Wilson, M.: Superconducting Magnets. Oxford Science Publications (1983)
10. Sharma, N., Bedford, R.: Hysteresis Machines, Mumbai, India (2003)
11. Tsuboi, Y., et al.: Torque Characteristics of a Motor Using Bulk Superconductors in the Rotor in Transient Phase. IEEE Trans. Appl. Supercond. 13, 2210 (2002)
12. Inácio, D., Pina, J., Gonçalves, A., Ventim-Neves, M., Leão-Rodrigues, A.: Numerical and Experimental Comparison of Electromechanical Properties and Efficiency of HTS and Ferromagnetic Hysteresis Motors. In: 8th European Conference on Applied Superconductivity (EUCAS 2007), Brussels, Belgium (2007)
13. Ventim-Neves, M.: Máquina Assíncrona. Notes of Electrotécnica Teórica, Internal publication, FCT-UNL

Appendix

A1 - Power Output and Torque Analysis: HTS Hysteresis Motor

According to [9], and since the applied magnetic field is equal to the full penetration field [10], the AC losses in the HTS materials are proportional to the hysteresis loop and frequency, as

$$P_{AC} = P_{H/cicle} \cdot f_{mag}, \quad (3)$$

where $P_{H/cicle}$ is given approximately by ξH_{appl}^2 and ξ is a factor depending on the characteristics and geometry of the superconductor. Assuming that the applied load torque is less than the motor torque, the motor accelerates until it reaches the synchronism, during the sub-synchronous regime the frequency of the rotor's magnetization is given by $f_{ro} = p(f_{ss} - f_{mec}) = p.f_{ss}.s$. f_{ss} is the mechanical synchronous speed, in revolutions per second (if f_{sup} is the supply's frequency, in Hz, then $f_s = f_{sup}/p$). To each value of the frequency f , there is a corresponding angular speed $\omega = 2\pi f$. The frequency f_{ro} is the frequency of the induced currents in the

rotor – in this case, in the HTS material – and therefore is the magnetization frequency in (3). The AC losses power are, therefore, given by

$$P_{AC} = P_{H/cycle} \cdot p \cdot f_s \cdot s, \quad (4)$$

The mechanical power, for the energy conservation principle, neglecting power and electric losses, results

$$P_{mec} = P_{elect} - P_{H/cycle} \cdot p \cdot f_s \cdot s. \quad (5)$$

At the motor starting, the slip is 1 and the mechanical speed is zero, so $P_{mec} = 0$. Replacing in (6) gives

$$P_{elect} = P_{H/cycle} \cdot p \cdot f_s \Rightarrow P_{mec} = P_{elect} (1 - s). \quad (6)$$

From (6) it is possible to observe that when the rotor accelerates from the start until the synchronous speed, the hysteresis power losses decreases, the developed mechanical power increases and the electrical power is kept constant.

The mechanical torque, T_{elmec} , is given by

$$T_{elmec} = \frac{P_{mec}}{\Omega_{mec}} \Rightarrow T_{elmec} = \frac{p^2}{2\pi} P_{H/cycle}. \quad (7)$$

With the mechanical speed given by

$$\Omega_{mec} = \omega_s \cdot (1 - s) = \frac{2\pi \cdot f_{sup}}{p} \cdot (1 - s) \quad (8)$$

A2 – Notations

R	Resistance
X	Reactance
λ	leakage reactance
subscript s	Stator
subscripts r	Rotor
subscripts m	Mutual/magnetization
subscripts p	Iron
subscripts c	Load
Indice ‘	Referred to stator
P	Power per phase
subscript $elect$	Electric
subscript AC	Alternating current
subscript mec	Mechanical
subscript $H/cycle$	hysteresis losses per cycle

H	Magnetic field
subscript $appl$	Applied
f	Frequency
subscript mag	Magnetization
subscript ss	Synchronous speed
subscript sup	Supply
subscript mec	Rotor's mechanical speed
subscript ro	Rotor's induced currents
T	torque
subscript $elmec$	Electromechanical
p	Number of poles pair
Ω_{mec}	mechanical speed

Transverse Flux Permanent Magnet Generator for Ocean Wave Energy Conversion

José Lima, Anabela Pronto, and Mário Ventim Neves

Faculdade de Ciências e Tecnologia – Universidade Nova de Lisboa,

Quinta da Torre, 2829-516 Caparica, Portugal

jose.a.o.lima@gmail.com, amgl@fct.unl.pt, ventim@uninova.pt

Abstract. Modern energy demands led the scientific community to renewable energy sources, such as ocean wave energy. The present work describes a model for a cost efficient rotary electrical generator, optimized for ocean wave energy conversion. The electrical power, supplied by low speed mechanical movement, requires the use of electrical machinery capable of generating high amounts of torque. Among the analyzed topologies, the one selected for further study was the Transverse Flux Permanent Magnet machine (TFPM). This topology differs from the conventional ones, presenting high power and torque densities, and allowing to independently set machine current and magnetic loadings in the machine. The machine was designed and analyzed through the use of a 3D FEM software. The obtained results show that the TFPM is a strong candidate to be used in large scale converting systems.

Keywords: Transverse flux, TFPM, Ocean wave energy, Low Speed, Generator, Finite elements.

1 Introduction

The ocean presents itself as an inexhaustible source of clean and renewable energy that appears mainly in the form of ocean waves, generated by the action of wind on the ocean surface, and in the form of ocean currents, caused by the effect of tides and by variations of salinity and temperature. Nowadays, there's a growing worldwide demand for energy resources that reveal themselves as alternatives to the existing ones, highly pollutant and with limited availability. Therefore, conditions must be created for their exploitation in a sustainable manner. Since the beginning of research on wave energy, encouraged by the oil crisis [1], several devices have been proposed to exploit this resource, although only a small number of these have been studied and implemented on a large scale [2]. Due to the complexity of ocean waves' characteristics to extract energy, the development of technologies that may take advantage of this energy source requires further research.

A vast knowledge about the physical aspects of ocean wave energy already exists [3] [4]. However, there is still no consensus on the best technology to take advantage of this resource.

The current work attempts a qualitative research on the main topologies of electrical machines that may allow a direct and efficient exploitation of low speed

rotational movement, and the selection, sizing and optimization of a topology for a small-scale electrical generator prototype. For each topology, the various pros and cons were considered. The Transverse Flux Permanent Magnet machine with flux concentrators was selected as the object of study, regarding its favorable characteristics for ocean wave energy conversion.

2 Contribution to Sustainability

Sustainability of Earth's energy resources urgently demands the optimization of its use, due to the current energy requirements and consumption. This work takes a further step towards turning ocean wave energy into a viable and desirable energy source. The developed generator's characteristics may act as an incentive to the expansion of ocean wave energy conversion.

3 Selected Electrical Generator

3.1 Direct Drive Approach

The ocean's environment is variable and unstable. A system with the purpose of converting mechanical ocean wave energy into electrical energy must be prepared to generate, with relatively high quality and efficiency, energy that meets the functional requirements of the electrical grid. Given the slow and undulatory motion of ocean waves, and to avoid the use of expensive gearboxes with periodic maintenance requirements, the development of an efficient direct drive electric generator was one of the main goals of this work. Thus, it becomes necessary to employ a high torque density machine, and therefore with a high number of poles.

3.2 Transverse Flux Permanent Magnet (TFPM) Topology

Several electrical generator topologies were analyzed. The TFPM topology seems to have great potential for ocean wave energy conversion, showing better use of the machine size and permanent magnet materials when compared to conventional and other non-conventional topologies [5]. Besides the possibility of reaching a very high torque density, by increasing the number of poles, it allows to set the electric current density regardless of the magnetic flux, unlike conventional radial topology machines, where the cross sectional area of air gap competes directly with the windings for the available space [6] [7] [8] [9]. In TFPM the air gap area defines the current density, while the axial length defines the magnetic flux density.

Due to its characteristics, the TFPM, with flux concentrators, was the topology selected for this research work.

4 Sizing

As aforementioned, the main objective of this work was to develop a model for a rotating electrical generator capable of operating at low speeds, allowing the

conversion of ocean wave mechanical energy into electrical energy to be supplied to the main grid. A single-phase TFPMP prototype model capable of generating an output of 10 kW was designed. It was considered an average angular frequency of 150 rpm for a gearless energy conversion system, at the shaft that drives the generator. Thus, in order to comply with the grid's frequency (50 Hz), the machine was designed with 20 pole pairs. Optimal values for each parameter of the generator topology were calculated and simplified machine's schemes were used to obtain expressions that reflect the generator's operation conditions.

4.1 Magnetic Circuit

The studied topology is illustrated in Fig. 1. The generator's rotor consists in two rows of permanent magnets and flux concentrators, and a central stainless steel separator, positioned between each row of the rotor, which acts as a "magnetic insulator" since it is made of a diamagnetic material. Each of these rows contains a set of magnets polarized in the direction of rotation; each magnet is inversely polarized with its pair in the opposite row and is also followed by a flux concentrator. The permanent magnet material chosen for this analysis was the Neodymium Iron Boron (NdFeB) due to its known magnetic properties [10], with a remanence $B_R=1.4T$ and a coercive field $H_C = 795 \text{ kA} \cdot \text{m}^{-1}$.

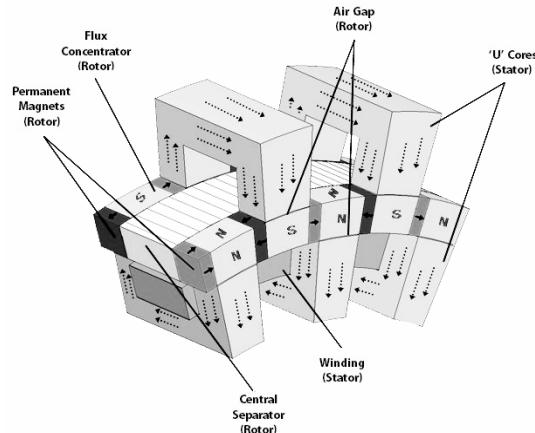


Fig. 1. TFPMP with double stator and single winding

The flux concentrator is made of electrical steel due to its high magnetic permeability, in order to aggregate a great amount of magnetic flux, reducing the possible reluctance and leakage flux of the circuit. In this double stator topology poles in 'U' are displaced in each stator and separated from the rotor by two air gaps for each pole. At the stator, the copper winding is displaced. In this type of topology a double winding may be used. This work makes an initial analysis with one winding displaced at the lower stator due to its simpler and robust construction, as shown in Fig. 1. For optimization purposes a deeper analysis with double winding topology is studied later.

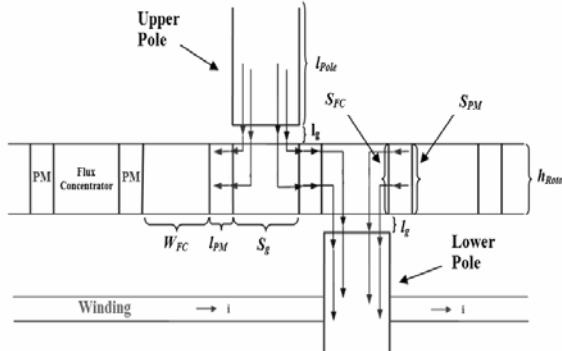


Fig. 2. Magnetic circuit: longitudinal view

Part of the magnetic flux path is presented in Fig. 2, as well as some of the dimensioned machine's parameters. To ease the machine's design and sizing, the following assumptions were made:

- The stator pole length l_{Pole} , is the same for the upper and lower pole;
- The section of both rows of permanent magnets of the rotor's flux concentrators is square and is expressed by $S_{\text{Rotor}} = h_{\text{Rotor}}^2$;
- Both permanent magnet and flux concentrator sections have the same value $S_{\text{PM}} = S_{\text{FC}} = S_{\text{Rotor}}$.

4.2 Working Point

Reaching a working point that corresponds to an efficient use of the materials and machine dimensions was one of the guidelines of this work.

Through the analysis of the topology's magnetic circuit depicted in Fig. 2, it was observed that the permanent magnet's maximum energy product demands a permanent magnet two times thicker than the air gap distance ($l_{\text{PM}} = 2l_g$). This would result in exceedingly long air gaps or extremely thin permanent magnets. To compensate such effect, using a relationship between the permanent magnet and the air gap sections given by a constant $K = S_g / S_{\text{PM}}$, it is possible to achieve an energy product somewhat closer to the material's maximum, allowing adequate air gap and permanent magnet dimensions. From Maxwell's equation that translates Ampere's law, and using Gauss's law for magnetism, the load line expression for the previously described magnetic circuit is given by:

$$\frac{B_{\text{mp}}}{H_{\text{mp}}} = -K \cdot \frac{\mu_0 \cdot l_{\text{mp}} \cdot \mu_{\text{fc}} \cdot \mu_p}{(h_{\text{rotor}} + w_{\text{fc}}) \cdot \mu_p + 2 \cdot l_g \cdot \mu_{\text{fc}} \cdot \mu_p + K \cdot l_p \cdot \mu_{\text{fc}}} \quad (1)$$

Fig. 3 shows the behavior of the maximum value for permanent magnet and 'U' pole magnetic flux density, depending on the constant K .

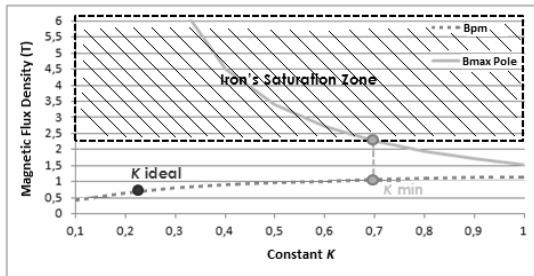


Fig. 3. Magnetic flux density with the variation of K

Due to poles shape, and in order to avoid its magnetic saturation, the ideal value for permanent magnets working point could not be reached. Therefore, the next best possible value for K was used. Nevertheless, with some pole shape enhancement, this characteristic can be improved.

For the described sizing a rotor's thickness h_{Rotor} of 5 cm and an air gap of 1mm were assumed, resulting in a machine's rotor radius of 31 cm.

4.3 Induced EMF

In the TFPMP topology each flux path is shared by two magnetic circuits as represented in Fig. 2. These circuits can be modeled by the electrical circuit showed in Fig. 4.

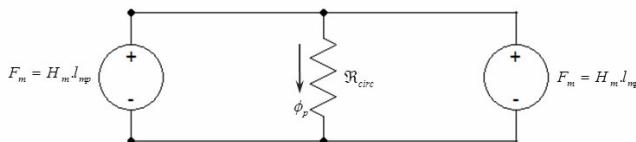


Fig. 4. Representative electrical scheme of one pole pair flux path

As the MMF sources are both equal and in parallel, defining P_{Poles} as the number of pole pairs, the induced EMF can be given by:

$$e(t) = N \cdot P_{\text{Poles}} \cdot \phi_p \cdot \omega \cdot \sin(\omega \cdot t) \quad (2)$$

5 Simulations

In order to verify the analytical expressions concerning machine's sizing and output characteristics, a graphical model of the topology was built and several simulations were performed, through the use of a 3D finite element method (FEM) software. Each of machine's design characteristics was parameterized, allowing an easier analysis of the machine's behavior.

Fig. 5 shows the magnetic flux density on each point of the constructed model and the flux path along a pole pair, respectively. The registered values matched the working point calculations made at the design phase, with no significant flux leakage observed.

The generator was tested with a purely resistive load sized for a demanding value equal to the nominal electrical power, delivering 8 kW, a value close to the expected. The flux linkage and the respective induced EMF and electrical current curves are shown in Fig. 6.

For each key parameter, various sets of tests were made, varying its value through a specified interval. Fig. 7 (a) depicts the study of the permanent magnet thickness, while maintaining the rotor radius. It was observed that the magnetic saturation is reached for permanent magnets thicker than 8 mm. The chosen value of 1 cm for the machine's permanent magnet thickness was very close to the optimum value.

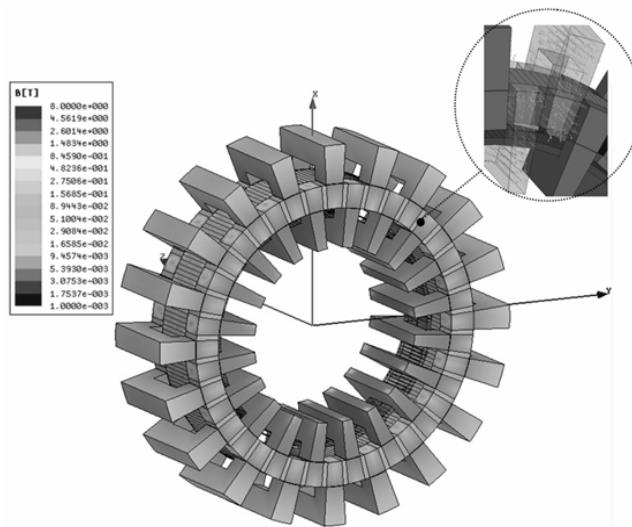


Fig. 5. FEM model's magnetic flux density and flux path along a pole pair

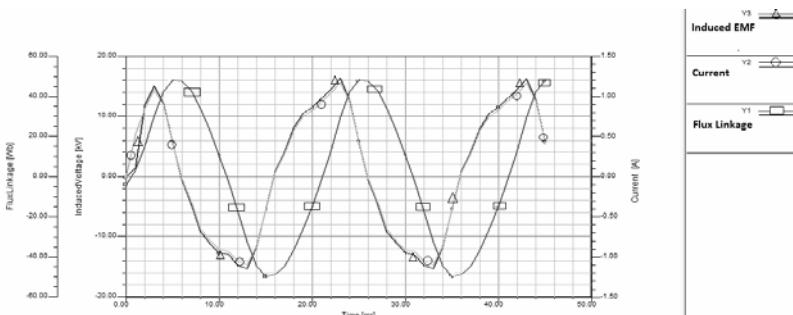


Fig. 6. Flux linkage, induced EMF and electrical current

For scalability purposes an analysis was made regarding the relationship between the torque density and the overall machine costs through a proportional increment of every machine parameter dimensions, except for the air gap length and the number of winding turns. The calculations were based on the specific costs used in [5]. The plot shown in Fig. 7 (b) reflects an improvement of efficiency in terms of the material's use. This characteristic may be an incentive to apply the TFPM generator in large scale systems.

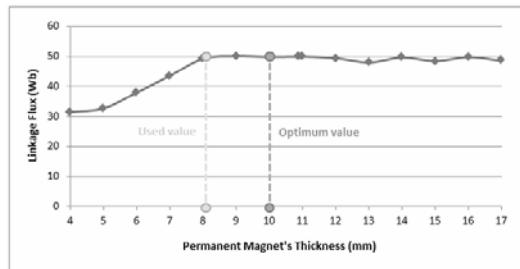


Fig. 7. Permanent magnet's thickness study

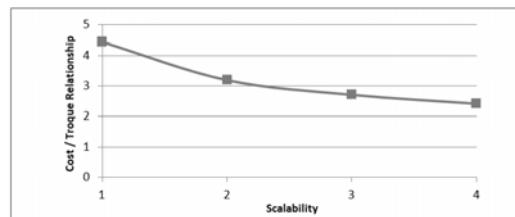


Fig. 8. Study of the scalability effects

6 Torque Density Optimization

The generator was sized and tested for a nominal power output of 10 kW, using a 3D FEM software. Although this goal was accomplished, the machine dimensions are underused. The same volume may be used more efficiently, generating more power and therefore improving the machine's torque density. Reinforcing the lower winding with conductors of larger diameter [11] and placing a second winding on the upper stator results in a better use of the permanent magnets material and machine dimensions. Through simulation, it was possible to obtain, with the same volume, an output power of 30,8 kW and a corresponding value of torque density $T_d = 22,5 \text{ kN} \cdot \text{m} \cdot \text{m}^{-3}$, which is close to the values shown in [5].

7 Conclusions

The TFPM machine proved to be a good alternative in direct drive ocean wave energy conversion, due to its topology and torque density characteristics, showing results

consistent with other similar studies and significant advantages over other topologies [5]. The reduction in the Cost/Torque relationship acts an incentive to the use of the TFPMP generator in full scale prototypes. As future work is intended a study of the topology under variable speed conditions and an optimization analysis of the 'U' poles shape for a working point improvement.

References

1. CEO, Potencial e estratégia de desenvolvimento da energia das ondas em Portugal. Centro de Energia das Ondas (2004)
2. Boud, R.: Status and Research and Development Priorities, Wave and Marine Accessed Energy, Dept. of Trade and Industry (DTI), DTI Report # FES-R-132, AEAT Report # AEAT/ENV/1054, United Kingdom (2003)
3. CRES: Ocean Energy Conversion in Europe, Centre for Renewable Energy Sources (2006)
4. Holthuijsen, L.H.: Waves in oceanic and coastal waters. Cambridge University Press, Cambridge (2007) ISBN 0521860288
5. Dubois, M.R., Polinder, H., Ferreira, J.A.: Comparison of generator topologies for direct-drive wind turbines. In: Proc. 2000 Nordic Countries Pow. and Indust. Elec., pp. 22–26 (2000)
6. Dubois, M.R.: Optimized Permanent Magnet Generator Topologies for Direct-Drive Wind Turbines. PhD thesis, Delft University of Technology, Delft, The Netherlands (2004)
7. Rang, Y., et al.: Analytical design and modelling of a transverse flux permanent magnet machines. In: IEEE PowerCon 2002, Kunming, China, October 13–17, pp. 2164–2167 (2002)
8. Arshad, W.M., Bäckström, T., Sadarangani, C.: Analytical design and analysis procedure of transverse flux machines. In: IEMDC 2001, pp. 115–121 (2001)
9. Lu, K.Y., Ritchie, E., Rasmussen, P.O., Sandholdt, P.: Modeling and power factor analysis of a single phase surface mounted permanent magnet transverse flux machine. In: Proc. 2003 IEEE Conf. Power Electronics and Drive Systems(PEDS), vol. 2, pp. 1609–1613 (2003)
10. Arnold Magnetic Technologies: Magnets Cathalogs,
<http://www.arnoldmagnetics.com>
11. PowerStream: Wire Gauge and Current Limits,
http://www.powerstream.com/Wire_Size.htm

A Fractional Power Disk Shaped Motor with Superconducting Armature

Gonçalo F. Luís¹, David Inácio², João Murta Pina², and Mário Ventim Neves²

¹ Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa
Monte de Caparica, 2829-516 Caparica, Portugal

² Centre of Technology and Systems,
Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa
Monte de Caparica, 2829-516 Caparica, Portugal
gfl19047@fct.unl.pt, ddi@fct.unl.pt, jmmp@fct.unl.pt,
ventim@uninova.pt

Abstract. A disk-shaped, double stator, induction motor with High Temperature Superconducting (HTS) field coils is proposed in this paper. Copper, typically used in windings of classic machines, limits current density allowed in field coils due to Joule effect losses. Also iron, which is used in magnetic circuits, limits the magnetic flux density obtained in the air gap due to saturation. The application of HTS field coils and iron removal effect in fractional power disk shaped or axial flux motors is analyzed by comparison of two different stator topologies. Twelve HTS field coils made of Bi-2223 ($\text{Bi}_2\text{Sr}_2\text{Ca}_2\text{Cu}_3\text{O}_{10}$) first generation tape, wrapped around a racetrack-shaped nylon core, are assembled. A simple topology was chosen, consisting of six filed coils per semi-stator arranged in the same plane with 60 ° displacement. This topology is analyzed theoretically, based on a linear induction motor approach and simulated using a commercial finite elements program, based on the same approach. In order to study the effect of magnetic saturation two stators were built. In the first, the field coils are assembled in steel plates. In the second, the same coils are assembled on nylon plates. The rotor is composed of an aluminum disk assembled on a stainless steel shaft. The HTS coils were cooled by liquid nitrogen (77 K). Simulations, experimental and theoretical results are consistent, showing high space harmonic distortion for the chosen topologies. It is shown that for this type of low power motors operating at this temperature, as iron saturation is not achieved, ferromagnetic materials removal is not a good option. Besides, flux leakage is to high, degrading developed torque.

Keywords: Superconducting tape, armature, induction, disk-shaped, ironless, space-harmonics.

1 Introduction

Due to High Temperature Superconducting (HTS) materials ability to carry high current densities with minimum losses when compared to conventional conductors, it is possible to design more compact and lightweight electric motors.

One of the advantages of superconducting conductors is their capability to create high intensity magnetic induction fields. Nevertheless, if ferromagnetic materials are used, due to saturation, flux density is limited to less than 2 T. In [1] - [3], some studies in the field of ironless HTS motors are performed. The goal of the work presented in this paper is to further study the application of air core HTS motors.

HTS tapes mechanical limitations, turns motors' constructions more complex than conventional motors. In [4] - [10] several topologies of HTS motors are presented.

However a direct comparison between iron and ironless cored motors is never accomplished. In this paper, a simple design is proposed, in order to create two identical HTS motors prototypes and compare the effects of removing ferromagnetic materials.

2 Contribution to Technological Innovation

One the main issues in electrical generator and motors is the high power losses due to the high resistivity of the conductors. In order to reduce power consumptions and create a sustainable power usage, more efficient electrical machines are necessary. The ability of HTS machines to work with minimum power losses make them a good substitution for conventional ones.

Other potential advantage in HTS motors is the possibility of iron removal, as mentioned, thus making lighter and compact machines.

These devices may also find potential application in naturally cryogenic environments, as is the case of double-shaded craters near Moon's magnetic poles, where temperatures are typically bellow 50 K. Thus, there is a current need to research and develop this kind of devices, where classical techniques do not always apply.

3 Motors' Design

Two prototypes motors are projected; one using a toothless ferromagnetic circuit, referred to as topology T1; and another, referred as T2, which is a ironless version of the first.

Both are disk shaped (or axial flux), double stator, two poles, induction motors. Twelve 20 turns identical HTS, racetrack shaped, field coils, are used, in order to create the travelling filed. These coils are made of Bi-2223 first generation tape. The coils were made considering the mechanical limitations of the tapes. The length of BSSCO tape for each coil is about 7 m. The field coils were tested at 77 K showing an average critical current of 88 A.

The prototypes design was limited by the coils dimensions and shape. The rotor consists on an aluminum disk assembled on a stainless steel shaft. The rotor thickness was optimized using the goodness factor of the machine. Based on the correction factors referred in [11], the expression of topology T1 goodness factor is obtained

$$Q(e) = \frac{\pi^2 f \mu_0 \sigma_{Al} \tau_p c_{bob} K_S}{2 p} \cdot \frac{e}{g_0(e) k_1(e) k_{sk}(e) (1 + k_p(e))} \quad (1)$$

where f is the frequency, c_{bob} the length of one coil, p the number of pairs of poles, g_0 the airgap, τ_p the pole pitch, K_S the Russel-Norsworthy correction-factor due to rotor's overhang, k_1 the large airgap's correction factor, k_{sk} the correction factor due to the skin effect in finite thickness rotor plates and k_p the correction factor due to saturation of armature plates iron. See [11] for correction factors k_1 , k_{sk} and k_p and [12] for K_S .

The goodness factor evolution as a function of the rotor thickness, for a frequency of 50 Hz, is shown in Fig. 1. Four curves are shown representing different slips of 60%, 70%, 89% and 90%. From that figure it can be conclude that the rotor thickness that optimizes the goodness factor is in the range of 3 to 3.5 mm.

A key project's goal was to ensure the dimensions similarity of both prototypes to ensure a reasonable comparison. T1 motor's final design is shown in Fig. 2. The prototype T2 is identical to T1, except that the plates are built of nylon instead of magnetic steel.

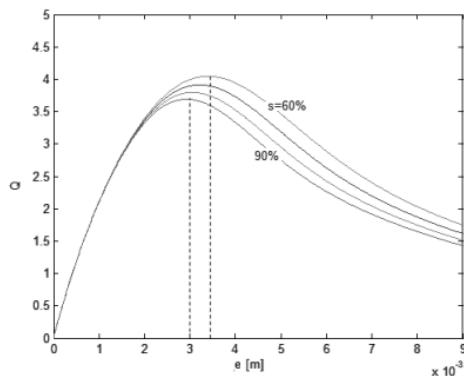


Fig. 1. Goodness factor as a function of rotor thickness, for different slips

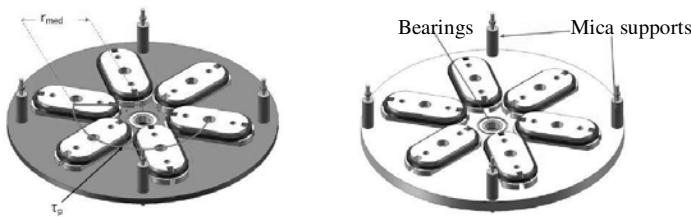


Fig. 2. Final aspect of T1 (left) and T2 (right) semi-stators design

4 Theoretical Analysis

In order to predict and analyze motors' operation, a theoretical study is performed. The main objectives are to observe the induction field created by the HTS coils and to predict the torque developed by the rotor. Only the prototype containing ferromagnetic circuits was studied, due to its simplicity.

4.1 Travelling Field

Based on Fourier analysis, the approximate expression of the travelling field, created by the armature, was calculated, showing great harmonic distortion. Fig. 3 shows the armature harmonics. From this figure it is possible to conclude that high order harmonics are not negligible. Thus, calculation of the developed torque is carried out considering the winding harmonics.

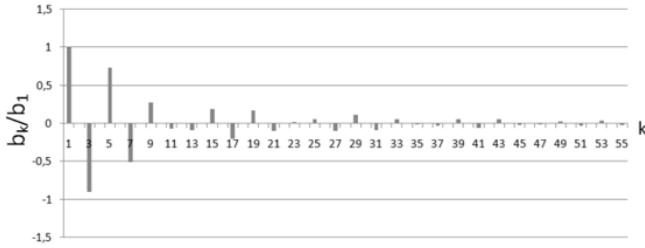


Fig. 3. Spectrum of winding harmonics, normalized by the fundamental's amplitude

4.2 Developed Torque

In order to determine the torque developed by the motor, considering the winding harmonics, the individual effect of each harmonic is calculated, and the total torque is the sum of each effect.

Based on [13], the expression of the torque developed by each component, T_k , is calculated as

$$T_k(\Omega) = r_{med} \pi J_1^2 b_k \cdot \frac{k \cdot \frac{e}{\rho_{vol}} \cdot (\omega_1 - \Omega k)}{\left(\frac{g}{\mu_0} \cdot k^2 \right)^2 + \left(\frac{e}{\rho_{Al}} \cdot (\omega_1 - \Omega k) \right)^2}, \quad (2)$$

where the rotor has a radius r_{med} , thickness e and ρ_{vol} is the volumetric resistivity of the aluminum. Also, J_1 is the current density in the HTS coils, b_k is the Fourier coefficient of the $k - th$ harmonic, ω_1 is the travelling field angular speed and Ω is the rotor's mechanical speed.

The individual components of the resultant torque are represented in Fig. 4.a), where it can be seen the effect of each harmonic k . The greater the value of the harmonic the greater the equivalent number of poles, therefore the smaller the equivalent synchronous speed.

Therefore the total developed torque is given by

$$T_T = \sum_{k=1}^{\infty} T_k(\Omega). \quad (3)$$

The evolution of torque T_T as a function of rotor's speed is plotted in Fig. 4.b), and it consists on the sum of the components shown in Fig. 4.a). It is clear that the function tends to a final speed of one third of the synchronous speed of a two pole motor, 3000 rpm. These values represent a slip of $s = 0,667$, considering no load operation. For a rotor's speed between 1000 and 2200 rpm, the motor behaves as a brake.

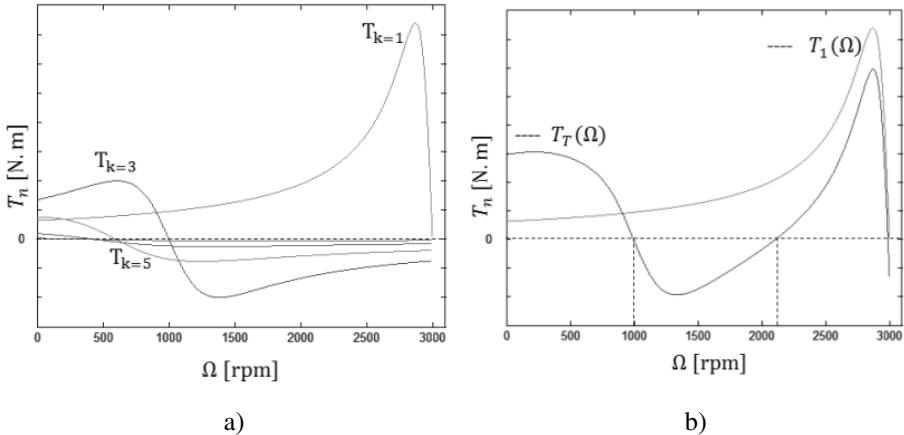


Fig. 4. a) Individual contribution of the main harmonics to the resulting torque. b) Resulting torque as a function of speed.

5 Simulations

Simulations were performed with three main goals. The first is to ensure the operation of the prototypes before building them. The second is to verify the theoretical results. The last is to study the effects of an ironless magnetic circuit and compare it with iron cored topologies.

The simulations were carried out using finite elements commercial software Flux2D, from Cedrat Company. Due to 2D geometry, a linearized version of the disk motor is studied. An electric current of 60 A per phase is used in simulations.

The T1 prototype is simulated and it is possible to verify the motor's operation. In order to verify the theoretical results, the torque/speed characteristic, $T_{T1}(\Omega)$, is obtained using imposed speed tests. Fig. 5 (T1) shows the obtained characteristic.

Comparing Fig. 5 (T1) and Fig. 4.b) it is possible to observe the similarities. In both graphics, the rotor's speed tends to a much lower value than the synchronous speed. Therefore, theoretical results correctly demonstrate the degradation of developed torque by winding harmonics.

Two other topologies were simulated. Topology T2, the ironless version of T1 and topology T3, which consists in adding more iron to the magnetic circuit. Fig. 5 presents torque/speed characteristic of the three topologies. It is possible to conclude that there is no advantage in removing the ferromagnetic circuits, besides devices'

weight reduction, since the torque decreases. Besides, since flux density only reaches 0.04 T in the airgap, it is possible to conclude that the magnetic field generated by the coils, is not enough to reach iron saturation, reinforcing previously conclusion.

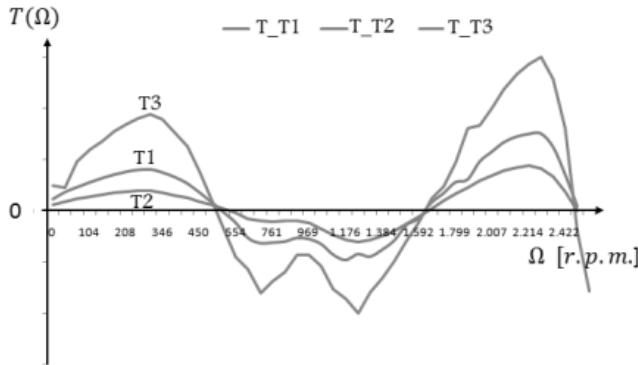


Fig. 5. Simulation results. Comparison of T1, T2 and T3 torque/rotor speed characteristics.

6 Experimental Results

The prototypes T1 and T2 were built accordingly with their designs, see Fig. 6. Both motors were tested at 77 K, with a 5 V power supply, with 63 A rms current supplying HTS coils. In order to obtain the torque/speed characteristic, the motors were coupled to a powder brake and several operation points were obtained. These are plotted in Fig. 7 for topology T1. For topology T2, the developed torque was neither enough to overcome nor static, nor dynamic friction. Therefore, no rotor movement is possible.

From these results, it is confirmed that there is no advantage in removing iron materials from the magnetic circuits, at least for this level of power. It is also possible to conclude that, as predicted in theoretical and simulations studies, the windings distribution on the stator results in a large slip.

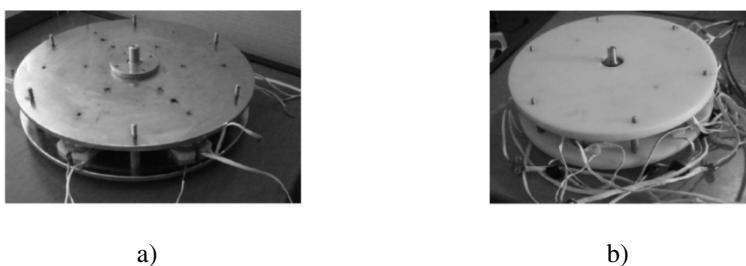


Fig. 6. Final look of the built prototypes: a) Topology T1. b) Topology T2.

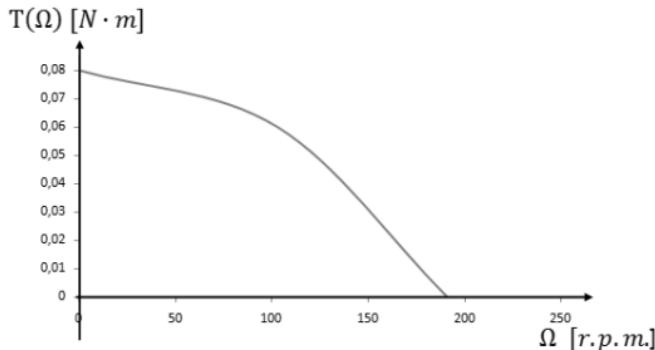


Fig. 7. Experimental torque/speed characteristic for topology T1

7 Conclusions

Two axial flux induction motor topologies were presented in this paper, both with HTS armature. Theoretical analysis, simulations and experimental measurements were performed, showing two main conclusions: firstly, there is no advantage in removing ferromagnetic circuits in this type of fractional power HTS motors, since the magnetic field created by the HTS tapes is not enough to compensate for the increased reluctance. Secondly, due to materials' physical restrictions, it was demonstrated that the windings distribution chosen, although simple, introduces high harmonic distortion resulting in poor motor performance. The HTS first generation tapes show great mechanical limitations, hindering the construction of more efficient motors. Also, a new approach for analytic study of the developed torque, for high winding harmonics motors, is presented. It shows results consistent with simulations and experimental results.

References

1. Masson, P.J., Luongo, C.A.: High Power Density Superconducting Motor for All-Electric Aircraft Propulsion. *IEEE Trans. Appl. Supercond.* 15(2) (2005)
2. Kwon, Y.K., Baik, S.K., Lee, E.Y., Lee, J.D., Kim, J.M., Kim, Y.C., Moon, T.S., Park, H.J., Kwon, W.S., Hong, J.P., Jo, Y.S., Ryu, K.S.: Status of HTS Motor Development for Industrial Applications at KERI & DOOSAN. *IEEE Trans. Appl. Supercond.* 17(2) (2007)
3. Granados, X., Pallares, J., Sena, S., Blanco, J.A., Lopez, J., Bosch, R., Obradors, X.: Ironless armature for high speed HTS disk shaped rotor in self levitating configuration. *Physica C* 372-376, 1520–1523 (2002)
4. Kovalev, L.K., Ilushin, K.V., Koneev, S.M., Kovalev, K.L., Penkin, V.T., Poltavets, V.N., Gawalek, W., Habersreuther, T., Oswald, B., Best, K.-J.: Hysteresis and Reluctance Electric Machines with Bulk HTS Rotor Elements. *IEEE Trans. Appl. Supercond.* 2(9), 1261–1264 (1999)
5. Nagaya, K., Suzuki, T., Takahashi, N., Kobayashi, H.: Permanent Magnet Pulse Motor With High-Temperature Superconducting Levitation. *IEEE Trans. Appl. Supercond.* 11(4), 4109–4115 (2001)

6. Hull, J.R., SenGupta, S., Gaines, J.R.: Trapped-Flux Internal- Dipole Superconducting Motor&Generator. *IEEE Trans. Appl. Supercond.* 9(2), 1229–1232 (1999)
7. Dombrovski, V., Driscoll, D., Shoykhet, B.A., Umans, S.D., Zevchek, J.K.: Design and Testing of a 1000-hp High-Temperature Superconducting Motor. *IEEE Trans. Energy Conversion* 20(3) (2005)
8. Masataka, I., Akira, T., Masayuki, K., Yoshiji, H., Toshihiro, S., Yoshihiro, I., Takashi, S., Yutaka, Y., Teruo, I., Yuu, S.: Development of a 15kW Motor with a Fixed YBCO Superconducting Field Winding. *IEEE Trans. Appl. Supercond.* (17), 1607–1610 (2007)
9. Schiferl, R., Flory, A., Livoti, W.C., Umans, S.D.: High Temperature Superconducting Synchronous Motors: Economic Issues for Industrial Application. In: PCIC 2006 Conference. IEEE, Los Alamitos (2006)
10. Pina, J.M., Neves, M.V., McCulloch, M.D., Rodrigues, A.L.: Design of a linear synchronous motor with high temperature superconductor materials in the armature and in the field excitation system. *Journal of Physics: Conference Series* 43, 804 (2006)
11. Boldea, I., Nasar, S.A.: Linear Motion Electromagnetic devices. Taylor & Francis, Abington (2001)
12. Rodrigues, A.L.: Design of Low Speed Linear induction Motor. M.Sc. thesis in Power Systems and Electrical Machines, Imperial College of Science and Technology (1973)
13. Yamamura, S.: Theory Of Linear Induction Motors, 2nd edn. University of Tokyo Press (1978)

Numerical Design Methodology for an All Superconducting Linear Synchronous Motor

João Murta Pina¹, Mário Ventim Neves¹, Alfredo Álvarez²,
and Amadeu Leão Rodrigues¹

¹ Centre of Technology and Systems,

Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa,
Monte de Caparica, 2829-516 Caparica, Portugal

jmmp@fct.unl.pt, ventim@uninova.pt, leao@uninova.pt

²“Benito Mahedero” Group of Electrical Applications of Superconductors,
Escuela de Ingenierías Industriales, University of Extremadura

Avenida de Elvas s/n, 06006 Badajoz, Spain

aalvarez@unex.es

Abstract. One potential advantage of the application of superconducting materials in electrical machines is the possibility to build lighter and compact devices by removing iron. These machines find applications, e.g., in systems where cryogenics is already available, or in naturally cryogenic environments. The design of motors with high temperature superconductors (HTS) presents issues unconsidered in classical machines, besides considerations on cryogenics, such as HTS brittleness or mechanical restrictions. Moreover, HTS’ electromagnetic properties also degrade due to flux density components, which arise if there is no iron to guide magnetic flux. Several aspects must thus be considered in the design stage, as applications may turn less attractive or even unfeasible. In this paper these issues are detailed, and a numerical methodology for the design of an all superconducting (without iron or conventional conductors) linear synchronous motor is presented.

Keywords: High temperature superconductivity, ironless motor, linear synchronous motor.

1 Introduction

The motives underlying the use of high temperature superconducting (HTS) materials in electrical motors are guided by their particular features that make them, in certain applications, advantageous when comparing to devices with conventional conductors and/or permanent magnets. Amongst these characteristics highlights the transport of high currents with minimum losses (up to 10^4 A/cm² at 77 K and 5 T, in second generation HTS tapes [1]) when compared to copper or aluminum conductors, the consequent generation of flux densities equivalently high, and the phenomenon of magnetic flux pinning. It is thus sometimes possible to remove iron from the devices, allowing for lighter and compact machines. HTS machines find potential application in environments where cryogenics is already available, as power industry [2-4] or

induction heating plants [5], or in naturally cryogenic environments, as future scientific stations in double-shaded craters near Moon's poles [6], where temperature does not rise above 50 K. The design of superconducting motors shows several specifications that do not arise in classical motors' design, besides considerations on cryogenics, such as HTS tapes bending limitations when these are used in armatures' windings. Besides, its electromagnetic properties also degrade due to the presence of flux density components perpendicular to tape surface, when there is no iron to guide magnetic flux. Consequently, applications may turn less attractive or even unfeasible.

Several types of HTS machines have been built in the past, as homopolar [7], synchronous [8-9], reluctance [10], hysteresis [11] and linear motors. Regarding the latter, HTS have been considered either in the armature [12] or excitation field [13].

This paper's main goal is to examine the key issues in the design of an all superconducting linear synchronous motor with HTS both in the armature and in the excitation system. The motor has no ferromagnetic materials or conventional electrical conductors, thus the designation "all superconducting". The final achievement is a numerical methodology for this type of motor design, incomparably faster than tools as e.g. finite elements software.

Next section examines this work's contribution to technological innovation. After that, the motor's architecture is described, followed by the description of the numerical methodology proposed for the determination of forces. Experimental results are shown in the sequel and conclusions are drawn in the end.

2 Contribution to Sustainability

HTS materials are foreseen as vehicles of important developments in the Energy field, allowing the advent of new technologies that would be unfeasible or even impossible with other approaches. Amongst some of the (many) current sustainability issues there is energy distribution in dense urban areas, integration of distributed generation in existing grids, or specific requirements to obtain lighter and compact electrical machines. For each of these problems, superconductivity has one or several answers, although commercial applications are still hard to find. Some factors contributing to this evidence are, besides the degree of reliability of current technologies, the need for cryogenics – whose trivialization depends too on the advent of HTS technologies – but also on the unavailability of practical and efficient design tools for HTS devices. This work intends to contribute to the latter issue, by presenting a methodology for aided design of HTS machines. One of this work's main achievements is that it allows replacing time consuming design tools like finite elements software, thus improving HTS technologies development, potential enablers of sustainable energy use.

3 Motor's Architecture

Ferromagnetic materials are used in electrical machines as guiders and amplifiers of magnetic flux. However, these materials impose restrictions in machines design: one related with magnetic induction saturation (typically bellow 1.8 T [14]), which leads to nonlinearities and complicates design; the other related with materials physical properties, namely its high density, that makes machines' weight and size determinant for high specific powers [15]. Moreover, hysteresis losses are generated in iron. These motivates the design of ironless motors, by the use of HTS materials.

3.1 Armature

The armature is built by single layer HTS tape Bi-2223 ($\text{Bi}_2\text{Sr}_2\text{Ca}_2\text{Cu}_3\text{O}_{14}$), with 90 A critical current, corresponding to 93 A/mm^2 engineering critical current density. This is one order higher than copper current density considered in machines design, about 4 A/mm^2 [14]. One armature's coil is represented in Fig. 1, and its parameters are described in Table 1. The windings were manufactured using previously machined nylon moulds. According to winding dimensions, the pole pitch, τ , is given by

$$\tau = 3(l_w + g) = 219 \text{ mm}. \quad (1)$$

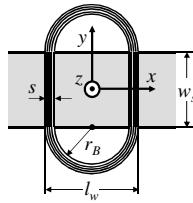


Fig. 1. Representation of one armature winding, build by Bi-2223 stacked tape

Table 1. Armature winding characteristics

Variable	Meaning	Value
N	Number of turns	20
s	Average coil's leg width	5 mm
r_b	Bending radius	30 mm
$l_w = 2 \times (r_b + s)$	Average coils width	70 mm
h_s	Coils height	4.2 mm
g	Average distance between adjacent coils	3 mm
w_s	Active coils' length	80 mm

The armature is built by a double stator, in order to minimize flux density components perpendicular to tape surface, i.e. in x direction, as was the case in this motor's previous geometries [16-18]. These components severely degrade tape's critical current.

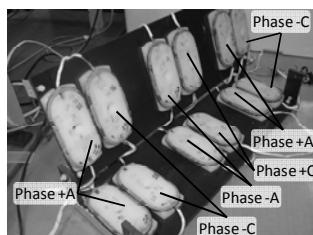


Fig. 2. Test armature built by a three-phase double stator, assembled for experimental measurements. Phase B is not implemented as is not necessary for static measurements, as later discussed.

A three-phase double stator test armature was built in order to make flux density measurements, see Fig. 2. Only phases A and B are shown, as later explained. Structural fastenings are built with nylon screws to avoid magnetic fields' distortions. In order to determine the developed static forces, the armature is considered to be fed by an ideal current inverter, see currents' profiles in Fig. 3.

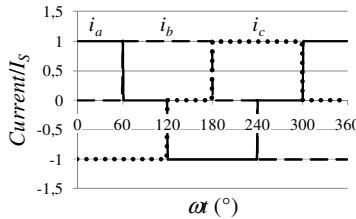


Fig. 3. Normalized armature's three-phase currents with amplitude I_S , generated by an ideal current inverter. i_a , i_b and i_c are currents from phases A, B and C, respectively.

3.2 Mover

The mover, comprising excitation, is built by two HTS Y-123 ($\text{Y}_{1.6}\text{Ba}_2\text{Cu}_3\text{O}_{7-x}$) bulks, with, see Fig. 4, magnetized prior to motor's operation. Sand-pile model [19] is used in order to numerically determine trapped flux, considering constant current according to Bean's model [20]. These already demonstrated results consistent with experiments [21]. Considering a $5.2507800 \text{ kA/cm}^2$ critical current, see later, and single domain bulks, the computed components of trapped flux, in a plane at 2 mm from bulk's surface, are shown in Fig. 5. The mover is represented in Fig. 6.

4 Numerical Determination of Developed Forces

In order to calculate the developed forces due to the interaction of armature currents with amplitude I_S and trapped fields, Laplace's law is applied to excitation,

$$d\vec{F} = -I_S d\vec{l} \times \vec{B}. \quad (2)$$

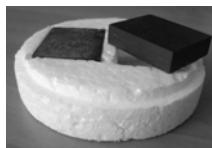


Fig. 4. Y-123 bulks (ATZ GmbH), used as trapped flux magnets for motor's excitation ($40 \times 32 \times 10 \text{ mm}^3$)

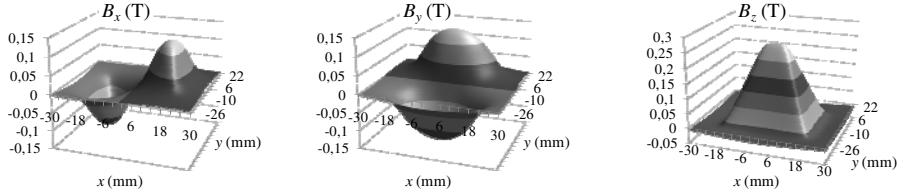


Fig. 5. Flux density components in an Y-123 bulk, numerically determined by sand-pile and Bean models, measured at 2 mm from bulk surface, and a critical current of 5.2507800 kA/cm²

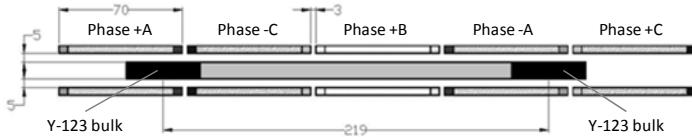


Fig. 6. Representation of the linear motor's mover between armature windings. All dimensions in millimeters.

This equation is integrated along the height and width of armature windings. According to the coordinate system defined in Fig. 1, and considering $d\vec{l} = -dy\hat{u}_y$, the several force components are derived, namely

$$\text{Thrust force: } dF_x = I_S B_z dy \quad (3)$$

$$\text{Vertical force: } dF_z = -I_S B_x dy \quad (4)$$

$$\text{Lateral force: } dF_y = 0. \quad (5)$$

Thus, B_y components do not contribute to developed forces, as they are parallel to armature current in the active region. First, the forces produced by a half stator are computed, and in the end the other half stator's contribution is added. Since flux density changes across windings height, its values are first averaged over z dimension, thus defining the following functions

$$B_x^{av}(x, y) = \frac{1}{h_s} \int_{h_s} B_x(x, y, z) dz \quad (6)$$

$$B_z^{av}(x, y) = \frac{1}{h_s} \int_{h_s} B_z(x, y, z) dz. \quad (7)$$

These functions are plotted in Fig. 7. After that, (6) and (7) must be averaged across winding's active length, thus originating functions

$$B_{xy}^{av}(x) = \frac{1}{w_s} \int_{w_s} B_x^{av}(x, y) dy \quad (8)$$

$$B_{zy}^{av}(x) = \frac{1}{h_s} \int_{h_s} B_z^{av}(x, y) dy, \quad (9)$$

which are plotted in Fig. 8. Functions B_{xy}^{av} and B_{zy}^{av} must contain the two Y-123 bulks, magnetized in opposite directions. The complete functions are plotted in Fig. 9.

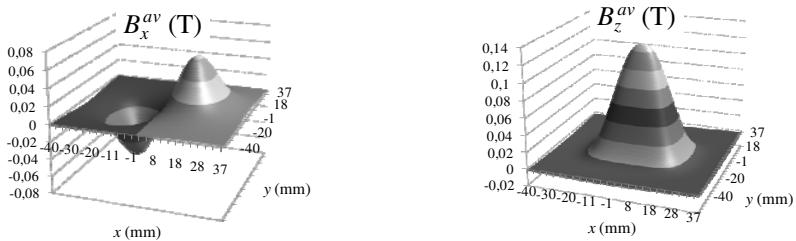


Fig. 7. Flux density average of one Y-123 trapped flux bulk along winding's height

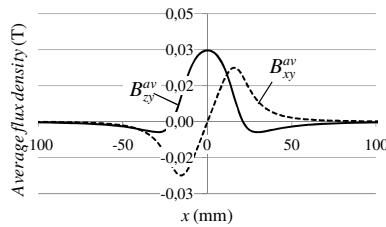


Fig. 8. Flux density average of one Y-123 trapped flux bulk along winding's height and length

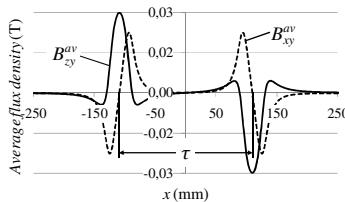


Fig. 9. Flux density average of the two Y-123 trapped flux bulks along winding's height and active length

The mover's position is varied from $x = 0$ to 2τ , (or from $\theta = 0^\circ$ to 360° , where $\theta = 180^\circ x/\tau$ is the angular displacement). In each position, forces are calculated by

$$F_x(\theta) = \int_{-\infty}^{\infty} c(x) B_{zy}^{av} \left(x - \tau \frac{\theta}{180^\circ} \right) dx \quad (10)$$

$$F_z(\theta) = \int_{-\infty}^{\infty} c(x) B_{xy}^{av} \left(x - \tau \frac{\theta}{180^\circ} \right) dx . \quad (11)$$

In (10) and (11), $c(x)$ represents current elements along armature. This is plotted in Fig. 10 as a function of θ , taking values $\pm I_S N/s$ where there is current density and zero in the other regions. For static forces' calculation a time instant in range $\omega t \in [0^\circ, 60^\circ]$ is chosen, corresponding to $i_a = I_S$, $i_b = 0$ and $i_c = -I_S$, with $I_S = 65$ A, see Fig. 3. This is why phase B is not implemented. Equations (6) to (11) are numerically integrated, returning thrust and vertical static forces corresponding to only one half stator. The other half's contribution means that that thrust is doubled, and vertical force is subtracted from itself, thus resulting in zero for all positions. Thrust force is represented in Fig. 11. It is clear that thrust force is highly oscillating, which is a consequence of the high space harmonics' content of magnetomotive force, due to Bi-2223 windings. This creates control issues outside the scope of this paper.

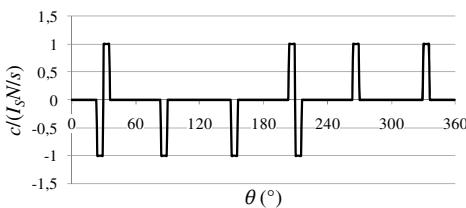


Fig. 10. Function $c(\theta)$ describing currents' profile along armature, for $\omega t \in [0^\circ, 60^\circ]$

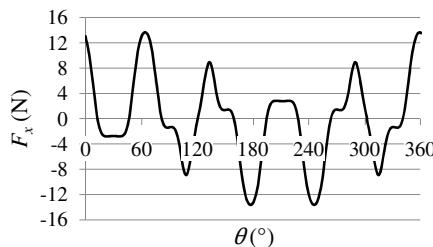


Fig. 11. Thrust force developed on the mover, as a consequence of the two half stators

4.1 Measurement of Flux Density Generated by the Armature

To measure flux density profile, a transversal flux Hall probe assembled in a xyz positioner was used, phases A and C fed by, respectively, 65 and -65 A, and phase B set to zero. The profile along a longitudinal path is plotted in Fig. 12, as well as the consistent results obtained by simulation with finite elements software Flux2D.

4.2 Measurement of Trapped Flux Density

In order to magnetize Y-123 bulks, four 400 A DC current sources in parallel were used. The bulk is placed in the middle of two air core inductors, supplied by a 1000 A current peak from the sources. Magnetic flux gets trapped in the bulk's pinning

centers, magnetizing it. A transversal Hall probe assembled on the same positioner was then used to measure trapped flux profile, shown in Fig. 13. Maximum flux density is 209 mT, which at 77 K cannot be increased, as the bulk is fully penetrated. By fitting curves obtained with sand-pile and Bean models, considering one single domain, it is possible to estimate a current density of 5,2507800 kA/cm². This is the value used in thrust force determination. The modeled profile is shown in Fig. 14.

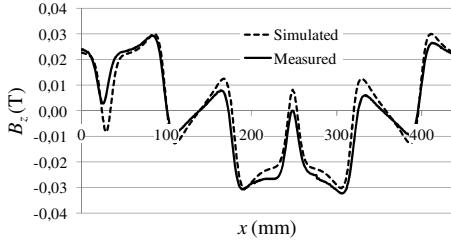


Fig. 12. Flux density component B_z measured along a longitudinal path across the armature, at middle distance between the two half stators. Simulated field is also shown for comparison.

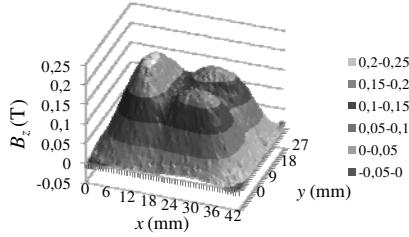


Fig. 13. Flux trapped in one bulk, after a current pulse. The peak from the left corresponds to one domain, while the two other peaks correspond to the other domain, which is damaged.

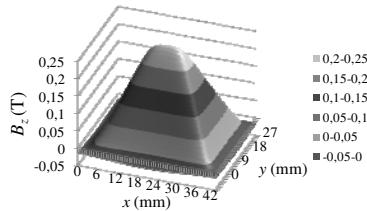


Fig. 14. Modeling of flux trapped in one Y-123 bulk, by sand-pile and Bean models

5 Conclusions and Future Work

An all superconducting linear synchronous motor was presented in this paper, as well as the methodology to derive the forces developed by this device. The main advantage of this motor is its reduced weight when compared with conventional motors, due to the absence of iron as well as copper conductors. This work's main goal, which is the

development of a numerical methodology that allows avoiding time consuming finite elements software for HTS devices design, is accomplished. Future work includes finishing the motor, validate numerical results, and compare its performance when iron is included.

Acknowledgments. This work was supported by FCT (CTS multiannual funding) through the PIDDAC Program funds.

References

1. Lee, P.J. (ed.): *Engineering Superconductivity*, pp. 260–280. Wiley-IEEE Press (2001)
2. Maguire, J.F., Yuan, J.: Status of high temperature superconductor cable and fault current limiter projects at American Superconductor. *Physica C* 469(15-20), 874–880 (2009)
3. Reis, C.T., Mehta, S.P., McConnell, B.W., Jones, R.H.: Development of High Temperature Superconducting Power Transformers. In: *Power Engineering Society Winter Meeting 2001*, pp. 432–437. IEEE, Ohio (2001)
4. Paul, W., Lakner, M., Rhyner, J., Unternährer, P., Baumann, T., et al.: Test of 1.2 MVA high-Tc superconducting fault current limiter. *Superc. Sci. Tech.* 10(12), 914–918 (1997)
5. Morandi, A., Fabbri, M., Ribani, P.L.: Design of a Superconducting Saddle Magnet for DC Induction Heating of Aluminum Billets. *IEEE Trans. Appl. Superc.* 18(2), 816–819 (2008)
6. Ryan, R.E., Underwood, L.W., McKellip, R., Brannon, D.P., Russel, K.J.: Exploiting Lunar Natural and Augmented Thermal Environments for Exploration and Research. In: *39th Lunar and Planetary Science Conference*, Texas (2008)
7. Superczynski, M.J., Waltman, D.J.: Homopolar Motor with High Temperature Superconductor Field Windings. *IEEE Trans. Appl. Superc.* 7(2), 513–518 (1997)
8. Frank, M., van Hasselt, P., Kummeth, P., Massek, P., Nick, W., et al.: High-Temperature Superconducting Rotating Machines for Ship Applications. *IEEE Trans. Appl. Superc.* 16(2), 1465–1468 (2006)
9. Buck, J., Hartman, B., Ricket, R., Gamble, B., MacDonald, T., Snitchler, G.: Factory Testing of a 36.5 MW High Temperature Superconducting Propulsion Motor. In: *Fuel Tank to Target: Building the Electric Fighting Ship* at American Society of Naval Engineers Day 2007, Arlington (2007)
10. Oswald, B., Best, K.-J., Setzer, M., Soll, M., Gawalek, W., et al.: Reluctance Motors with Bulk HTS Material. *Superc. Sci. Tech.* 18, 24–29 (2005)
11. Kovalev, L.K., Ilushin, K.V., Penkin, V.T., Kovalev, K.L., Koneev, S.M.-A., et al.: Hysteresis and Reluctance Electric Machines with Bulk HTS Elements. Recent Results and Future Development. *Superc. Sci. Tech.* 13(5), 498–502 (2002)
12. Kim, W.-S., Jung, S.-Y., Choi, H.-Y., Jung, H.-K., Kim, J.H., Hahn, S.-Y.: Development of a Superconducting Linear Synchronous Motor. *IEEE Trans. Appl. Superc.* 12(1), 842–845 (2002)
13. Sato, A., Ueda, H., Ishiyama, A.: Operational Characteristic of Linear Synchronous Actuator with Field-Cooled HTS Bulk Secondary. *IEEE Trans. Appl. Superc.* 15(2), 2234–2237 (2005)
14. Say, M.G.: *The Performance and Design of Alternating Current Machines: Transformers, Three-Phase Induction Motors and Synchronous Machines*, 3rd edn. CBS Publishers & Distributors, New Delhi (1983)

15. Vajda, I., Szalay, A., Gobl, N., Meerovich, V., Sokolovsky, V.: Requirements for the industrial application of superconducting rotating electrical machines. *IEEE Trans. Appl. Superc.* 9(2), 1225–1228 (1999)
16. Pina, J.M., Ventim Neves, M., McCulloch, M.D., Rodrigues, A.L.: Design of a linear synchronous motor with high temperature superconductor materials in the armature and in the field excitation system. *J. Phys: Conf. Series*, 43(1), 804–808 (2006)
17. Pina, J.M., Ventim Neves, M., Rodrigues, A.L.: Case Study in the Design of HTS Machines: an All Superconducting Linear Synchronous Motor. In: International Conference on Power Engineering, Energy and Electrical Drives, POWERENG 2007, Setúbal, pp. 185–190 (2007)
18. Pina, J., Gonçalves, A., Pereira, P., Álvarez, A., Ventim Neves, M., Rodrigues, A.: A test rig for thrust force measurement of an all HTS linear synchronous motor. *J. Phys: Conf. Series*, 97(1), 12220 (2008)
19. Fukai, H., Tomita, M., Murakami, M., Nagatomo, T.: Numerical simulation of trapped magnetic field for bulk superconductor. *Physica C* 357-360, Part 1, 774–776 (2001)
20. Bean, C.P.: Magnetization of High-Field Superconductors. *Rev. Mod. Phys.* 36, 31–39 (1964)
21. Aydiner, A., Yanmaz, E.: Numerical calculation of trapped magnetic field for square and cylindrical superconductors. *Superc. Sci. Tech.* 18(7), 1010–1015 (2005)

CMOS Fully Differential Feedforward-Regulated Folded Cascode Amplifier

Edinei Santin, Michael Figueiredo, João Goes, and Luís B. Oliveira

Departamento de Engenharia Electrotécnica / CTS – UNINOVA
Faculdade de Ciências e Tecnologia (FCT) / Universidade Nova de Lisboa (UNL)
2829-517 Caparica – Portugal
[{e.santin,l.oliveira}](mailto:{e.santin,l.oliveira}@fct.unl.pt)@fct.unl.pt, [{mf,jg}](mailto:{mf,jg}@uninova.pt)@uninova.pt

Abstract. A fully differential self-biased inverter-based folded cascode amplifier which uses the feedforward-regulated cascode principle is presented. A detailed small-signal analysis covering both the differential-mode and the common-mode paths of the amplifier is provided. Based on these theoretical results a design is given and transistor level simulations validate the theoretical study and also demonstrate the efficiency and usefulness of the proposed amplifier.

Keywords: fully differential amplifiers, feedforward-regulated cascode technique, self-biasing, inverter-based, CMOS analog integrated circuits.

1 Introduction

Amplifiers are essential building blocks used frequently to build feedback networks able to perform a variety of accurate functions, e.g. multiplication, addition, integration, inversion, etc. The accuracy of these operations is directly dependent on the amplifier's gain-bandwidth product (GBW) performance [1].

Dictated by the down scaling of the digital circuits, the CMOS technology evolved posing several obstacles to the analog circuits design in general and in particular to the amplifiers design. Some of these obstacles are low intrinsic gain (g_m/g_{ds}) of transistors, reduced supply voltages, high variability of devices, etc., which inevitably deteriorate the performance of the well-known amplifier topologies. To cope with this problem some existing amplifier topologies have been enhanced and novel topologies have been proposed, some recent examples are [2]-[4].

In this paper we propose a fully differential self-biased inverter-based folded cascode amplifier which uses the feedforward-regulated cascode principle firstly presented in [4]. First, the small-signal behavior of the topology is analyzed in detail. After, a design is outlined and transistor level simulations are presented to demonstrate the efficiency of the proposed new topology. Finally, the main conclusions are drawn.

2 Contribution to Sustainability

A new self-biased inverter-based transconductance amplifier topology using feedforward-regulated cascode devices is presented. The combination of these features

allows the topology to achieve an attractive figure of merit ($\text{FoM} = \text{GBW} \cdot C_L / \text{Power}$), which is comparable or better to those of the best amplifiers up to date. This enhanced bandwidth to power dissipation efficiency is interesting for high speed and low power applications to be realized in deep-submicron CMOS technology nodes ($\leq 0.13\text{-}\mu\text{m}$).

3 Amplifier Description and Analysis

3.1 Circuit Description

The circuit of the proposed amplifier without common-mode (CM) feedback (CMFB) circuitry is shown in Fig. 1. The input stage is composed of transistors M_1 , M_4 , M_{2a-b} , and M_{3a-b} , where M_{2a-b} and M_{3a-b} are responsible for converting input voltage variations in incremental drain currents which are then “folded” to the output cascode devices. These currents are converted to incremental voltages in nodes v_{D2a-b} and v_{D3a-b} , and these voltages are used to drive cascode transistors M_{6a-b} and M_{7a-b} in a feedforward fashion as proposed in [4]. For this reason the topology is named feedforward-regulated folded cascode. An interesting advantage of the feedforward-regulated cascode topology over the well-known feedback-regulated cascode one is the faster cascode regulation, which leads to a higher operation speed [4]. The inverter-based inputs of the amplifier effectively double the input g_m compared to single-transistor inputs, which can be used favorably in attaining high GBW.

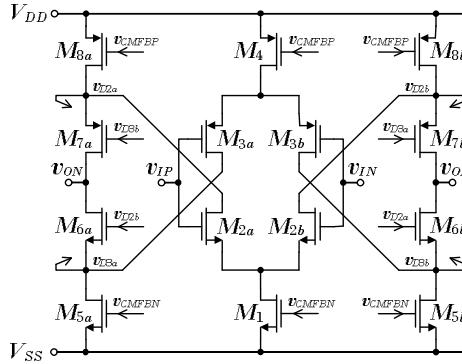


Fig. 1. Electrical schematic of the proposed amplifier (CMFB circuitry not shown)

The extra (self) biasing voltages v_{CMFBN} and v_{CMFBP} , which are also used to adjust the output CM voltage, are obtained from the circuit depicted in Fig. 2. This circuit performs two functions: 1) it averages the output voltage obtaining $v_{CMFB} = (v_{OP} + v_{ON})/2 \equiv v_{CMO}$; 2) in the sequence, it level-shifts this voltage down, resulting in v_{CMFBN} , and up, resulting in v_{CMFBP} . These voltage level-shifts are needed to bias transistors M_1 , M_4 , M_{5a-b} and M_{8a-b} in the saturation region without sacrificing considerably the output voltage swing. For example, assuming $V_{DD} = 1.2$ V, $V_{SS} = 0$ V, and $V_{CMO} = 0.6$ V, the $|V_{DS}|$ (drain-source voltage) of the aforesaid transistors should be greater than 0.3 V to saturate these transistors in a technology with a $|V_{TH}| \approx 0.3$ V (threshold

voltage). As a result, 50 % of the rail-to-rail voltage is consumed, which is not desirable. Regarding the averaging circuit, the value of the resistors should be sufficiently large in order not to lower the output impedance, and hence the gain. Unfortunately, large resistor values mean large silicon area, and, if this is prohibitive, other CMFB circuitry, depending on the application, can be used (e.g., switched-capacitor CMFB, differential difference amplifier CMFB, etc.). It is important to remark that besides the rail voltages (V_{DD} and V_{SS}) no additional biasing voltages are required, i.e., the amplifier is completely self-biased. This feature precludes the use of an explicit biasing circuit, saving power consumption and silicon area.

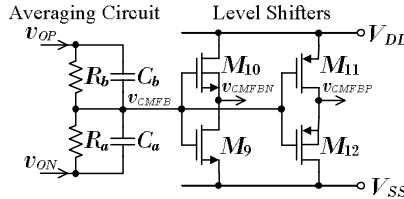


Fig. 2. Electrical schematic of the CMFB circuitry

3.2 Circuit Analysis

The small-signal differential-mode (DM) analysis is carried out with the equivalent half-circuit shown in Fig. 3 (a). Here C_L represents an output capacitive load. The small-signal model used for all MOS transistors is illustrated in Fig. 3 (b).

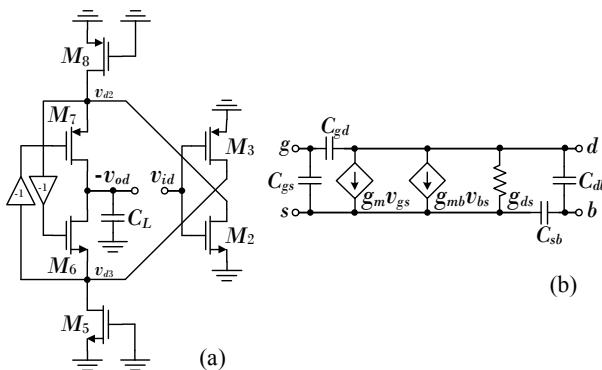


Fig. 3. Small-signal DM half-circuit of the amplifier (a) and MOS transistor model (b)

By considering a “perfect” symmetry, that is, the small-signal parameters of transistors M_2 , M_5 , and M_6 identical to those of M_3 , M_8 , and M_7 , respectively, the following input-output transfer function results:

$$\frac{V_{od}}{V_{id}} = \frac{2(sC_{gd2} - g_{m2})(sC_{gd6} - g_{eq1})}{C_{eq1}C_{eq2}s^2 + [C_{eq1}(g_{eq1} + g_{ds2} + g_{ds5}) + 2g_{ds6}(C_{eq2} + C_{gd6})]s + 2g_{ds6}(g_{ds2} + g_{ds5})} \quad (1)$$

where $g_{eq1} = 2g_{m6} + g_{mb6} + g_{ds6}$, $C_{eq1} = C_L + 2C_{gd6} + 2C_{db6}$, and $C_{eq2} = C_{gd2} + C_{db2} + C_{gd5} + C_{db5} + 2C_{gs6} + C_{sb6}$. It is important to mention that, besides the two poles and two zeros present in (1), a pole-zero doublet occurs between the dominant and nondominant poles. With the symmetry assumption, this doublet is perfectly canceled out. If we consider the dominant pole angular frequency ω_d much lower than that of the nondominant pole ω_{nd} , expressions for these poles can be derived from (1) as follows [5]:

$$\omega_d = \frac{2g_{ds6}(g_{ds2} + g_{ds5})}{C_{eq1}(g_{eq1} + g_{ds2} + g_{ds5}) + 2g_{ds6}(C_{eq2} + C_{gd6})} \approx \frac{2g_{ds6}(g_{ds2} + g_{ds5})}{C_{eq1}(g_{eq1} + g_{ds2} + g_{ds5})} \quad (2)$$

and

$$\omega_{nd} = \frac{C_{eq1}(g_{eq1} + g_{ds2} + g_{ds5}) + 2g_{ds6}(C_{eq2} + C_{gd6})}{C_{eq1}C_{eq2}} \approx \frac{g_{eq1} + g_{ds2} + g_{ds5}}{C_{eq2}}. \quad (3)$$

The low-frequency (DC) gain is given by:

$$A_{dc} = \left(1 + \frac{2g_{m6}}{g_{ds6}} + \frac{g_{mb6}}{g_{ds6}} \right) \left(\frac{g_{m2}}{g_{ds2} + g_{ds5}} \right). \quad (4)$$

It is important to note that bulk-source transconductances g_{mb} of transistors M_{6a-b} and M_{7a-b} increase the gain. Therefore, it is recommended not to eliminate the body-effect of these transistors, even though this is possible in most modern CMOS technologies. Also from (1) we see the transfer function has two right-half plane zeros. In practice, these zeros are at relatively high frequencies and can be neglected. The maximum GBW for this amplifier, considering a phase margin of about 65° , i.e. the nondominant pole located at a frequency twice the GBW [1], is $\omega_{nd}/(2^*2\pi)$ hertz.

We now analyze the small-signal common-mode behavior. For this purpose, we consider the equivalent circuit shown in Fig. 4, where it is assumed ideal averaging circuit as well ideal level shifters. One important requisite for the CM path is to have sufficient bandwidth (generally equal or greater than that of the DM path [6]) and good phase margin to not deteriorate the DM performance. It is also desirable a moderate DC gain to stabilize the output CM voltage with some accuracy.

By visual inspection, we see that the overall transfer function (V_{cmo}/V_{cmfb}) of the circuit in Fig. 4 is of fifth order, since it has five signal nodes, excluding the input. It is now clear why, for simplicity, we made ideal both the averaging circuit and the level shifters (otherwise the transfer function would be of seventh order and hardly manageable). Also, if desirable, the influence of these circuits can be included a posteriori. Considering symmetrical devices, i.e., transistors M_1 , M_2 , M_5 , and M_6 identical to M_4 , M_3 , M_8 , and M_7 , respectively, two perfectly matched doublets arise and the system reduces to “third” order. In this case the transfer function is given by:

$$\frac{V_{cmo}}{V_{cmfb}} \approx \frac{2(2C_{gd6}s + g_{eq2})\{2C_{gd5}C_{eq3}s^2 + [(C_{gd1} + 2C_{gd5})g_{eq1} + 2(C_{gd5}g_{ds1} - g_{m5}C_{eq3})]s - 2g_{m5}(g_{eq1} + g_{ds1}) - g_{m1}g_{eq1}\}}{C_{eq1}C_{eq2}C_{eq3}s^3 + [(g_{eq2} + 2g_{ds2} + 2g_{ds5})C_{eq3} + (g_{eq1} + g_{ds1})C_{eq2}]C_{eq1}s^2 + (g_{eq2} + 2g_{ds5})(g_{eq1} + g_{ds1})C_{eq1}s + 8g_{ds6}[g_{ds5}(g_{eq1} + g_{ds1}) + g_{ds1}g_{ds2}]} , \quad (5)$$

where $g_{eq1} = 2(g_{m2} + g_{mb2} + g_{ds2})$, $g_{eq2} = 2(g_{mb6} + g_{ds6})$, $C_{eq1} = 2C_L + 4C_{db6}$, $C_{eq2} = 2(C_{gd2} + C_{db2} + C_{gd5} + C_{db5} + C_{gd6} + C_{sb6})$, and $C_{eq3} = C_{gd1} + C_{db1} + 2(C_{gs2} + C_{sb2})$. This transfer function is approximated because the sum of products terms multiplying the complex frequency “ s ” powers in the denominator are somewhat simplified; however, each term has a simplification error no greater than 1 % for typical parameter values.

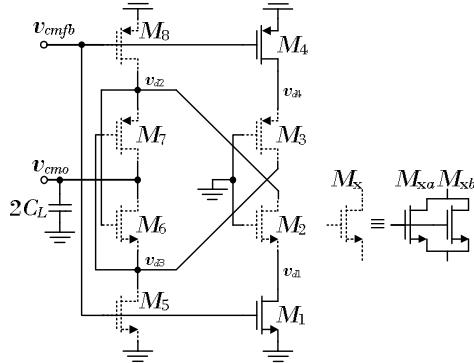


Fig. 4. Small-signal CM equivalent circuit considering ideal averaging circuit and level shifters

Considering a dominant-pole behavior, i.e. the frequency of the first pole much lower than the remaining ones, the dominant pole can be derived from (5) as:

$$\omega_{d,cm} = \frac{8g_{ds6}[g_{ds5}(g_{eq1} + g_{ds1}) + g_{ds1}g_{ds2}]}{(g_{eq2} + 2g_{ds5})(g_{eq1} + g_{ds1})C_{eq1}} \approx \frac{8g_{ds6}g_{ds5}}{(g_{eq2} + 2g_{ds5})C_{eq1}} . \quad (6)$$

Since the nondominant (second) pole is not far away from the third pole, it can not be derived using an approach similar to that used for the dominant one. Therefore, a different approach [7] is used to approximate this pole (with an error lower than 10 %) resulting in:

$$\omega_{nd,cm} = \frac{\frac{1.6(g_{eq2} + 2g_{ds5})(g_{eq1} + g_{ds1})}{(g_{eq2} + 2g_{ds2} + 2g_{ds5})C_{eq3} + (g_{eq1} + g_{ds1})C_{eq2}}}{\frac{0.045[(g_{eq2} + 2g_{ds2} + 2g_{ds5})C_{eq3} + (g_{eq1} + g_{ds1})C_{eq2}]}{C_{eq2}C_{eq3}}} . \quad (7)$$

Similar to the DM path, this nondominant pole will dictate the maximum CM GBW for a given stability (phase margin). Hence, expressions (3) and (7) play a crucial role in the analysis and design of the proposed amplifier. The CM DC gain is straightforwardly derived from (5) and is given by:

$$A_{dc,cm} = \frac{-g_{eq2}[2g_{m5}(g_{eq1} + g_{ds1}) + g_{m1}g_{eq1}]}{4g_{ds6}[g_{ds5}(g_{eq1} + g_{ds1}) + g_{ds1}g_{ds2}]} . \quad (8)$$

The CM transfer function (equation (5)) also has three zeros; however, these zeros are at relatively high frequencies and can be ignored in practice.

4 Amplifier Design

The accuracy of the feedback networks (settling error, gain accuracy, bandwidth, etc.) where amplifiers are usually used is directly related with the amplifier's loop gain, which in turn at a particular frequency is a function of the GBW; therefore, the greater the GBW can be designed, the better system performances can be obtained [1]. As a result, in a basic view the design (sizing) problem of an amplifier is to achieve the maximum GBW possible with the minimum average power dissipation. In practice, however, other requirements (e.g., output swing, input-referred noise, distortion, etc.) have to be considered simultaneously resulting in a very challenging problem. To deal with this complexity and be able to obtain the most efficient solution (sizing), it is a common procedure to use a circuit optimizer.

Alternatively to the automated/optimized sizing, a manual and interactive equation-based approach can be followed. In order to come up with a tractable design problem, only the most important requirements are considered by the designer and the remaining ones are achieved with some design rules of thumb, which generally mean over-sizing the circuit. Although this approach generally precludes the optimum solution, we follow it here to gain more insight into the amplifier operation and also to validate the usefulness of the relatively simple expressions derived in the analysis section.

Throughout the previous section we conveniently considered symmetrical devices. Unfortunately, the PMOS and NMOS transistors are not symmetrical. At a simplified view, they differ mostly in the effective mobility (in the technology used in this work the mobility ratio of the NMOS/PMOS transistors (α_{np}) is roughly four times). To compensate this lower PMOS performance, the aspect ratio (W/L) of the PMOS transistors is usually α_{np} times greater than that of the NMOS counterparts. With this constraint, both device types have the same g_m (for a given drain current I_D), and this is also good to improve the noise performance of the amplifier [1]. However, the price to pay for wider PMOS transistors is greater parasitic capacitances. Another important constraint is to use channel lengths two or more times greater than the minimum channel length (L_{min}) allowed by the technology. Doing so short-channel phenomena are highly attenuated, but also the devices are slowed down.

Table 1. Active devices sizing and biasing ($V_{OV} \equiv V_{GS} - V_{TH}$)

Devices	W/L [μm/μm]	I_D [μA]	V_{OV} [mV]	V_{DS}/V_{DSsat} [V/V]	g_m [mS]	g_{mb} [μS]	g_{ds} [μS]	C_{gs} [fF]	C_{gd} [fF]	C_{sb} [fF]	C_{db} [fF]
M_1	32/0.24	344	79	2.14	5.30	679	212	54	9	17	15
M_{2a-b}	24/0.24	172	42	7.66	3.10	341	71	37	7	11	8
M_{3a-b}	96/0.24	-177	-75	5.02	2.90	522	69	155	28	68	57
M_4	128/0.24	-354	-106	1.65	4.93	996	212	226	38	101	94
M_{5a-b}	24/0.24	285	79	4.44	4.36	554	114	41	7	13	10
M_{6a-b}	16/0.24	108	52	1.58	1.88	184	121	25	5	6	6
M_{7a-b}	64/0.24	-108	-73	2.53	1.79	305	49	103	19	42	40
M_{8a-b}	96/0.24	-280	-108	2.57	3.90	788	108	169	28	76	67
M_9	0.18/0.72	8	383	1.12	0.03	4	3	1	0.1	0.1	0.1
M_{10}	24/0.18	8	-111	18.07	0.20	26	6	8	7	13	8
M_{11}	1.20/0.72	-8	-335	1.32	0.04	8	1	7	0.4	1	1
M_{12}	67/0.18	-8	71	16.97	0.19	38	7	26	21	54	38

Following the previous constraints and the set of equations derived in the analysis section, the amplifier is sized targeting the highest GBW while providing a moderate DM DC gain (> 55 dB) and an average current consumption of about 1 mA. The sizing result is summarized in Table 1 (the passive elements sizing is: $C_L = 4$ pF, $C_{a-b} = 0.5$ pF, and $R_{a-b} = 500$ k Ω). It is worth to note that only transistors M_{10} and M_{12} are not saturated in strong or moderate inversion, and, except for the parasitic capacitances, the transistors small-signal parameters are approximately symmetrical.

5 Simulation Results

In this section some electrical simulations are presented for the amplifier sizing shown in Table 1, which is done in a standard 0.13- μ m high-speed 1.2 V CMOS technology ($L_{min} = 0.12$ μ m) from a pure foundry player. Only standard- V_{TH} devices are used. For all simulations a load capacitance of 4 pF is employed. The DM frequency response is depicted in Fig. 5. The amplifier achieves a GBW of 197 MHz with a phase margin of about 83°, and a DC gain of about 55 dB while dissipating 1.12 mW from a 1.2 V supply. This results in a FoM of 704 MHz·pF/mW, which compares favorably with other up to date amplifiers [8].

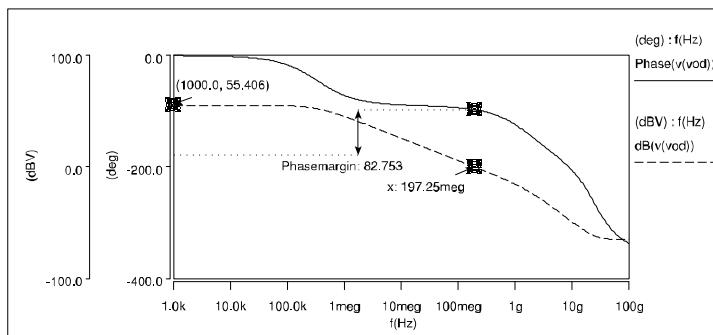


Fig. 5. Differential-mode frequency response

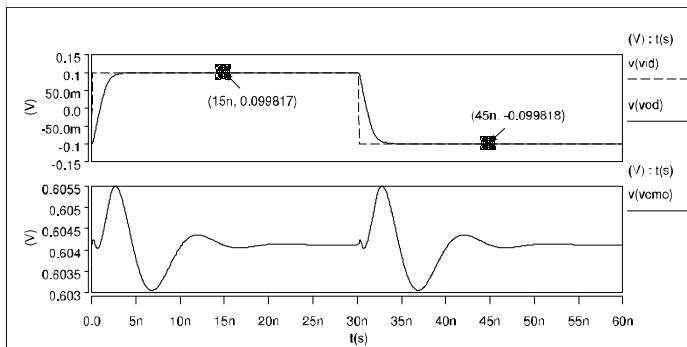


Fig. 6. 0.2 Vp-p,diff step response (top) and output CM voltage settling (bottom)

The step response of the amplifier in unity-gain feedback configuration for a step of 0.2 V_{p-p,diff} is shown in Fig. 6 (top), where the settling final values are clearly indicated. Also in this figure is shown the output CM voltage settling, which indicates a good CM behavior. The amplifier settles to within the 0.01 % error band about the final values in less than 7.1 ns (exactly 6.96 ns for the positive step and 7.06 ns for the negative one).

6 Conclusions

In this work we proposed a fully differential self-biased inverter-based folded cascode amplifier which uses the feedforward-regulated cascode technique. The amplifier achieves a DM GBW of 197 MHz with a capacitive load of 4 pF while dissipating an average power of only 1.12 mW from a 1.2 V supply. Configured as a unity-gain follower, and with the same load, the amplifier settles to within an error of 0.01 % in less than 7.1 ns for 0.2 V_{p-p,diff} positive and negative steps. These results show the suitability of the proposed amplifier for high speed and low power applications designed in standard deep-submicron CMOS technologies.

Acknowledgments. This work has been supported by the Portuguese Foundation for Science and Technology through projects SPEED (PTDC/EEA-ELC/66857/2006), LEADER (PTDC/EEA-ELC/69791/2006), IMPACT(PTDC/EEA-ELC/101421/2008) and TARDE (PTDC/EEA-ELC/65710/2006), and Ph.D. grants BD/62568/2009 and BD/41524/2007. The authors also would like to thank Flávio Gil for his initial contributions and Prof. Adolfo Steiger Garção for the many technical discussions which help to deepen the understanding of the proposed amplifier.

References

1. Steyaert, M., Sansen, W.: Opamp Design towards Maximum Gain-Bandwidth. In: Huijsing, J., van der Plassche, R., Sansen, W. (eds.) *Analog Circuit Design: Operational Amplifiers, Analog to Digital Convertors, Analog Computer Aided Design*, pp. 63–86. Kluwer Academic Publishers, The Netherlands (1993)
2. Assaad, R.S., Silva-Martinez, J.: The Recycling Folded Cascode: A General Enhancement of the Folded Cascode Amplifier. *IEEE J. Solid-State Circuits* 44, 2535–2542 (2009)
3. Figueiredo, M., Santin, E., Goes, J., Tavares, R., Evans, G.: Two-Stage Fully-Differential Inverter-based Self-Biased CMOS Amplifier with High Efficiency. In: Proc. IEEE Int. Symp. Circuits Syst (ISCAS), pp. 2828–2831 (May 2010)
4. Zheng, Y., Saavedra, C.E.: Feedforward-Regulated Cascode OTA for Gigahertz Applications. *IEEE Trans. Circuits Syst. I, Reg. Papers* 55, 3373–3382 (2008)
5. Razavi, B.: *Design of Analog CMOS Integrated Circuits*, pp. 173–177. McGraw-Hill, New York (2001)
6. Sansen, W.: *Analog Design Essentials*, pp. 239–262. Springer, The Netherlands (2006)
7. Palumbo, G., Pennisi, S.: *Feedback Amplifiers: Theory and Design*, pp. 248–249. Kluwer Academic Publishers, Dordrecht (2002)
8. Perez, A.P., Nithin Kumar, Y.B., Bonizzoni, E., Maloberti, F.: Slew-Rate and Gain Enhancement in Two Stage Operational Amplifiers. In: Proc. IEEE Int. Symp. Circuits Syst (ISCAS), pp. 2485–2488 (May 2009)

A New Modular Marx Derived Multilevel Converter

Luis Encarnaçāo¹, José Fernando Silva², Sónia F. Pinto², and Luis. M. Redondo¹

¹ Instituto Superior de Engenharia de Lisboa, Cie3, Portugal

luisrocha@deea.isel.pt, lmredondo@deea.isel.pt

² Instituto Superior Técnico, Cie3, TU Lisbon, Portugal

fernandos@alfa.ist.utl.pt, soniafp@ist.utl.pt

Abstract. A new Modular Marx Multilevel Converter, M^3C , is presented. The M^3C topology was developed based on the Marx Generator concept and can contribute to technological innovation for sustainability by enabling wind energy off-shore modular multilevel power switching converters with an arbitrary number of levels. This paper solves both the DC capacitor voltage balancing problem and modularity problems of multilevel converters, using a modified cell of a solid-state Marx modulator, previously developed by authors for high voltage pulsed power applications. The paper details the structure and operation of the M^3C modules, and their assembling to obtain multilevel converters. Sliding mode control is applied to a M^3C leg and the vector leading to automatic capacitor voltage equalization is chosen. Simulation results are presented to show the effectiveness of the proposed M^3C topology.

Keywords: Modular Multilevel, Capacitor voltage equalization, Marx modulator.

1 Introduction

Multilevel converters (MC) are the technology of choice for medium and high voltage flexible AC transmission systems (FACTS). Their industrial use is increasing in FACTS, as MCs enable the use of existing power semiconductors with nearly 5kV blocking capability to obtain converters able to operate at 100-300 kV. MCs are being preferred over conventional two-level converters, as the required high number of levels of their staircase output voltages additionally reduces total harmonic distortion (THD) and electromagnetic interference (EMI)[1].

However, well known MC topologies such as the Neutral-Point Clamped (NPC), flying capacitor (FC), and cascaded H-bridge (CHB), have strong limitations in balancing the DC capacitor voltage dividers that limit the semiconductor voltages to a few kV, when the required number of level increases beyond five. Some topologies such as NPC and FC are also not modular and their complexity increases with the square of the number of the levels required.

To solve these problems, half bridge based modular approaches (M^2LC) were proposed in 2001 [2]. However, the half bridge concept needs redundancy and must

sample all the capacitor voltages for the central processor to decide which power semiconductors should be switched on or off [3, 4, 5].

This paper solves the modularity problems of MCs and the DC capacitor voltage balancing, using a modified cell of the solid state Marx modulator, previously developed by authors for high voltage pulsed power applications [6]. The DC capacitor voltage measuring circuits and control complexity are completely avoided since the modified cell, called Modular Marx MC (M^3C), performs DC capacitor voltage balancing automatically, using just an extra switch without needing no DC capacitor voltage measurements.

After the Contribution to Sustainability (section 2), the paper details the structure and operation of M^3C modules and the assembling of basic cells to obtain MCs. Three and five-level M^3C topologies are presented (section 3), detailing the capacitor balancing in the three-level topology. Simulation results are presented in Section 4 to show the effectiveness of the proposed sliding mode controlled M^3C 5 level topology for two selected applications.

2 Contribution to Sustainability

In the emerging area of modular MCs, this paper proposes a new modular semiconductor cell, M^3C , to build high voltage high number of levels MCs for FACTS or DC-AC converters for off-shore wind parks. The Marx derived M^3C cell solves two main problems in MCs: 1) All M^3C cells are identical (modularity) and 2) they provide inherent balancing capability of all DC capacitor voltages avoiding voltage measuring circuits and regulation costs. Power Converters using M^3C cells will contribute to energy availability, regulation and cleanliness, enhancing energy sustainability.

3 Modular Multilevel Marx Converter M^3C

The M^3C modules and their assembling to obtain MCs are described. Circuit configurations for three-level and a five-level inverter legs are presented.

3.1 M^3C Cell and Three-Level MC Leg

The M^3C cell topology adds an extra switch, S_{EK} , (Fig. 1a), to each Marx basic cell [6], providing a bi-directional switch with the existing diode D_{EK} . Therefore, the charge of C_K capacitors in adjacent cells (Fig. 1b) can be equalized turning on switch pairs S_K , D_K and S_{EK} , D_{EK} .

The three-level MC leg topology uses two basic cells (Fig. 1b) for each half arm, with a total of 4 cells.

From (Fig. 1b), considering voltages $U_{CA} \approx U_{CB}$ and $U_{dc} = U_{CA} + U_{CB}$, each capacitor will be charged with voltage $U_{Ci} = U_{dc}/(n-1)$, where n represents the number of levels (in this case $n=3$, implying $U_{Ci}=U_{dc}/2$).

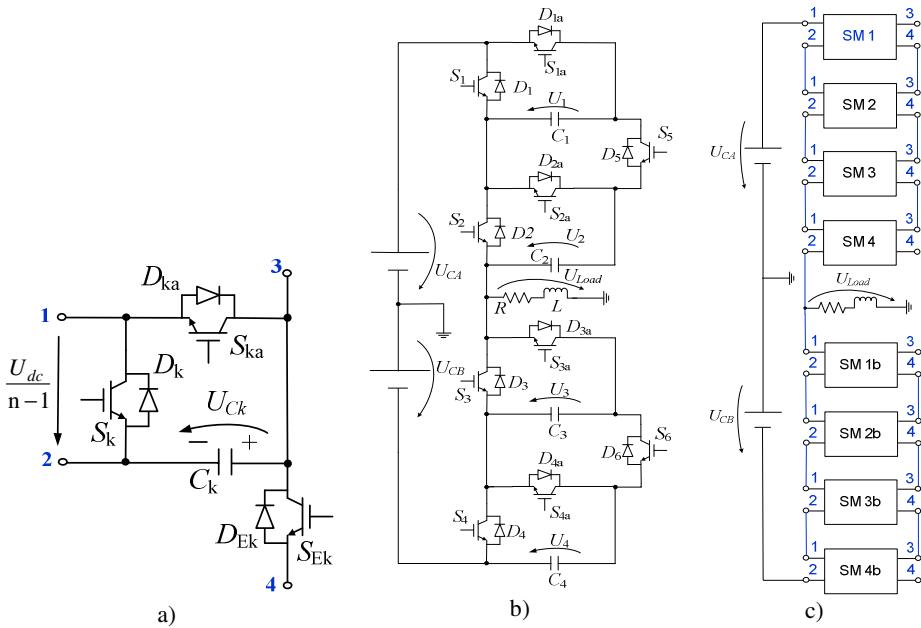


Fig. 1. Modular Multilevel Marx Converter topology: a) Structure of the basic cell; b) Three-Level M³C leg; c) Five-Level M³C leg using 8 cells

To understand the operating principles of three-level Modular Multilevel Marx Converter (Fig. 1b), Table 1 shows the three voltage levels of voltage U_{Load} and the number of turned on (S_K on) basic cells which are necessary to obtain those voltage levels (or voltage vectors) on each arm. Also, the number of possible redundant states for each level (vector) is shown.

Table 1. Voltage levels and number of vectors for a Three-Level M³C leg

Vector	U_{Load}	Number of ON Cells		Number of States = n possibilities upper Arm \times n possibilities bottom Arm
		Upper	Bottom	
1	$-U_{dc}/2$	0	2	$1 \times 1 = 1$
2	0	1	1	$2 \times 2 = 4$
3	$+U_{dc}/2$	2	0	$1 \times 1 = 1$

3.2 Five-Level M³C Leg

Using the basic M³C cell, n level MCs can be obtained, using $n-1$ basic cells for the upper arm, and $n-1$ cells for the bottom arm. Therefore, to obtain a five-level M³C eight basic cells are necessary for each converter arm (Fig. 1c). There are several redundant states in levels 2, 3, 4, depending on the state of each cell (Table 2).

Table 2. Voltage levels and number of vectors for a Five-Level M³C leg

Vector	U_{Load}	Number of ON Cells		Number of States
		Upper	Bottom	
1	$-U_{dc}/2$	0	4	$1 \times 1 = 1$
2	$-U_{dc}/4$	1	3	$4 \times 4 = 16$
3	0	2	2	$6 \times 6 = 36$
4	$+U_{dc}/4$	3	1	$4 \times 4 = 16$
5	$+U_{dc}/2$	4	0	$1 \times 1 = 1$

3.3 DC Capacitor Voltage Balancing

To illustrate the cell inherent balancing capability, consider for example, a three level leg with the two upper cells conducting (S_1 and S_2 on) to obtain $U_{Load} = U_{dc}/2$. Then the conduction of the extra switch (S_5 or D_5) parallels the two upper capacitors (C_1 , C_2) equalizing their charges. The equivalent happens in the bottom arm, with capacitors C_3 and C_4 , when applying the vector 1 to obtain the minimum level ($U_{Load} = -U_{dc}/2$).

Table 3 lists the switch states for all the operating vectors including the 4 possible states of vector 2 ($U_{Load} = 0V$). It is easy to see that the state V2a, in which S_{1A} , S_2 and S_5 conduct in the upper arm, also connects capacitors C_1 and C_2 in parallel equalizing their charge. Therefore this state should be the only one to be used for vector 2. Fig. 2 confirms the above reasoning by presenting f our simulation results ($U_{dc}=2000V$, $C_1=C_2=C_3=C_4=10\mu F$ and inductive load RL 1mH, 50Ω), each simulation using one state of vector 2. It is shown that using state V2a the capacitor voltages are balanced (Fig. 2a), while for remaining states (Fig. 2b, Fig. 2c and Fig. 2d), the capacitor voltages are unbalanced.

Table 3. States of semiconductors (1 if ON, 0 if OFF) for a Three-Level M³C leg

Level	State	S1	S1a	S2	S2a	S5	S3	S3a	S4	S4a	S6	U_{LOAD}
1	V1	0	1	0	1	0	1	0	1	0	1	$-U_{CB}$
2	V2a	0	1	1	0	1	0	1	1	0	1	0V
	V2b	1	0	0	1	0	1	0	0	1	0	
	V2c	0	1	1	0	1	1	0	0	1	0	
	V2d	1	0	0	1	0	0	1	1	0	1	
3	V3	1	0	1	0	1	0	1	0	1	0	U_{CA}
Upper Arm							Bottom Arm					

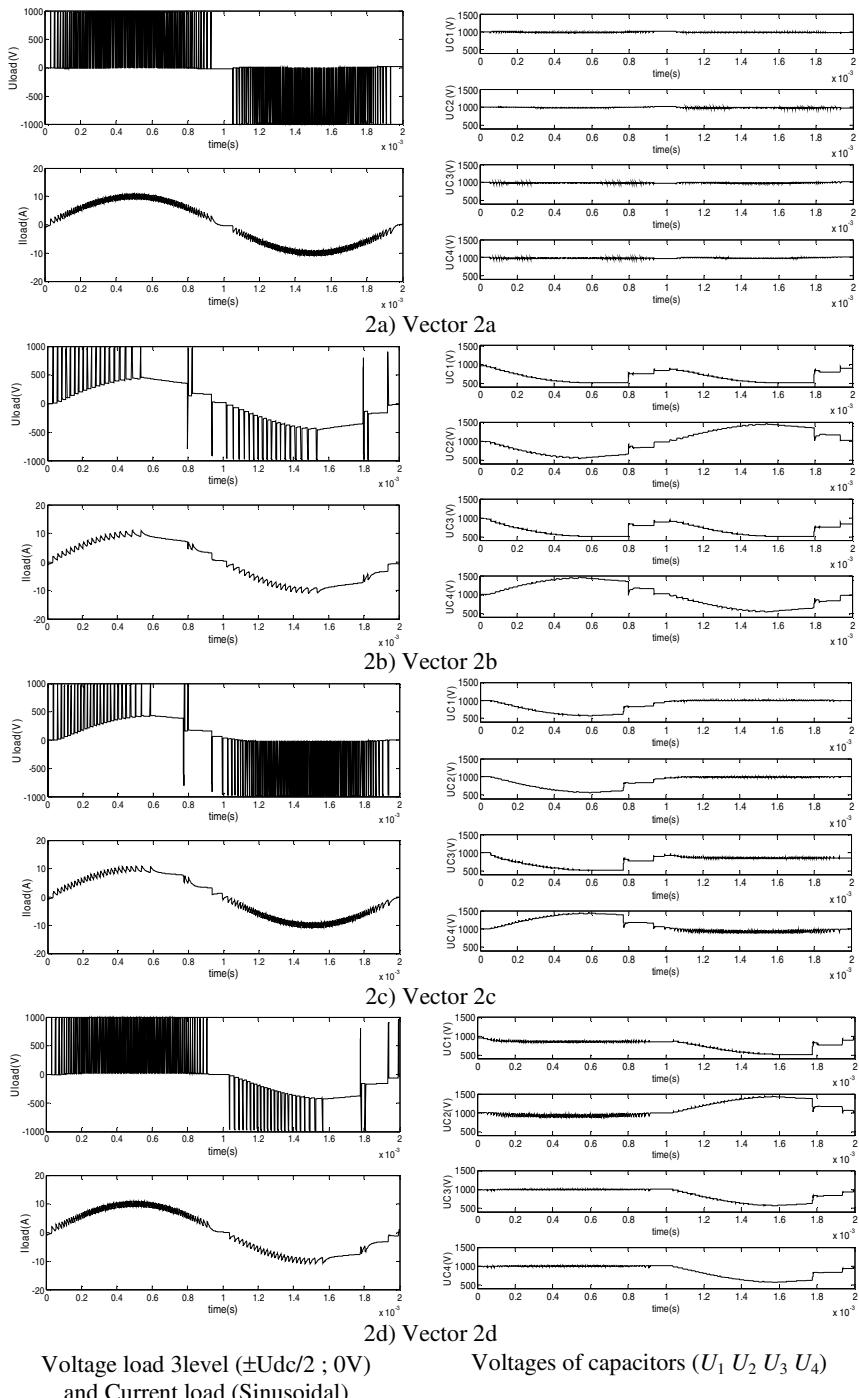


Fig. 2. Simulation results for the three-level arm obtained with vectors V2a, V2b V2c and V2d

4 Sliding Mode Controlled Five-Level M³C Leg

Two applications of five-level M³C are simulated in the Matlab/Simulink environment using a sliding-mode stability based multilevel modulator [7, 8, 9], according to Fig. 3 [7]. Circuit parameters are $U_{dc}=2000V$, $C_K =5\mu F$, $K_i=1000$ and capacitive load $R||C$ ($1M\Omega$, $100nF$) in series with $L=1nH$.

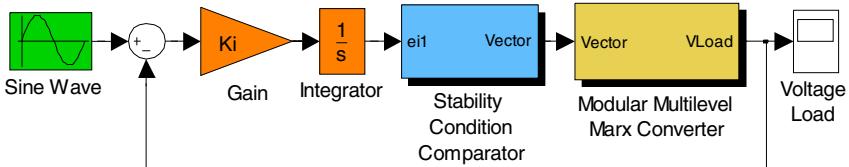


Fig. 3. Block diagram of the sliding-mode stability based multilevel modulator for M³C

The first application illustrates the M³C operating as a high voltage pulse generator (Marx Generator). The M³C was designed for five positive levels ($0V$; $\frac{1}{4} U_{dc}$; $\frac{1}{2} U_{dc}$; $\frac{3}{4} U_{dc}$, U_{dc}). The amplitude of the impulse reference is $1350V$. Sliding mode control is suitable to overcome the slow C_K capacitors discharge, usually called “voltage droop”. The sliding-mode stability based modulator ensures the desired voltage applied to the load by increasing or decreasing the chosen level (Fig. 4a) so that the mean value of the error of the controlled output voltage is near zero inside a tolerance band of $\pm 6mV$ (Fig. 4b). The M³C controller uses the third ($\frac{1}{2} U_{dc}=1000V$) and the fourth level ($\frac{3}{4} U_{dc} =1500V$) to maintain the desired output average value near $U=1350V$.

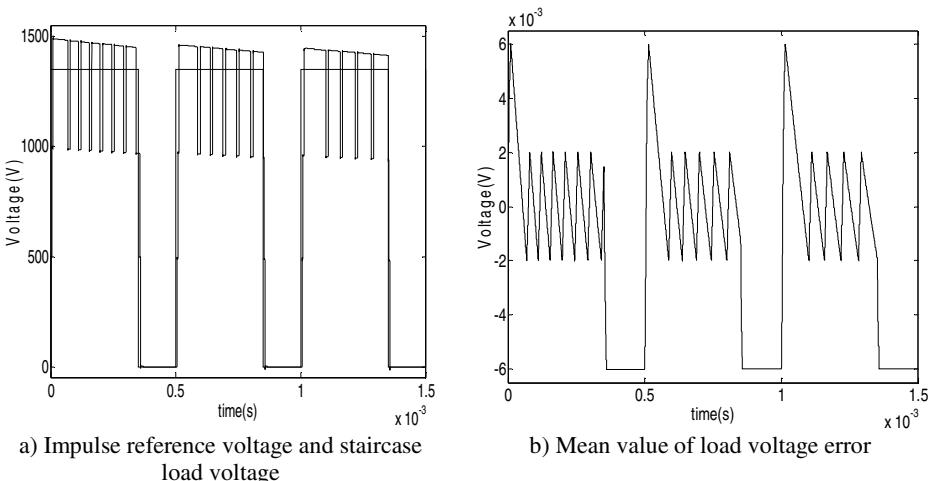
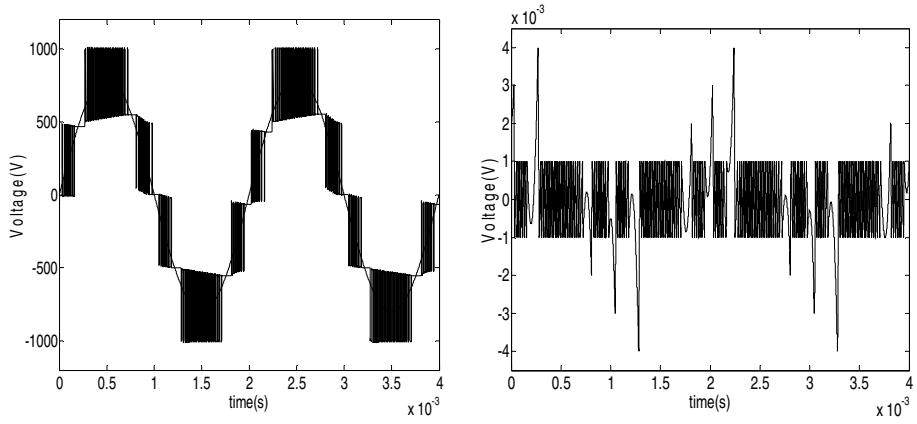


Fig. 4. Simulation results for M³C operating as a Marx Generator



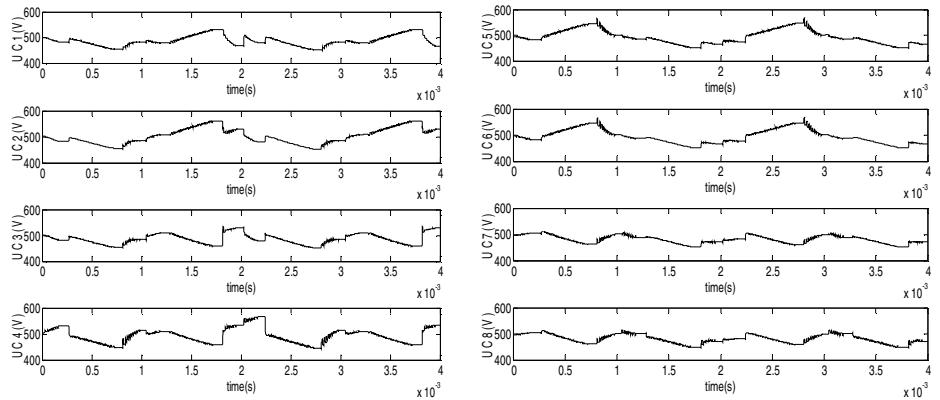
a) Sinusoidal reference voltage and staircase load voltage

b) Mean value of load voltage error

Fig. 5. Simulation results for M^3C operating as a five level inverter

In the second application, the 5 level M^3C operates as a multilevel inverter to deliver a sinusoidal output voltage with reference amplitude equal to 800V (Fig. 5a). In this case, the output voltage levels used are $\pm 1/2 U_{dc}$, $\pm 1/4 U_{dc}$, 0V. Fig. 5b presents the mean value of the error of the controlled output voltage showing it is nearly zero ($\pm 4\text{mV}$ tolerance).

Fig. 6 shows the 8 capacitor voltages obtained in this operation. The capacitor voltages are balanced within approximately $\pm 10\%$ of their working voltage.



6a) Upper Arm ($U_1 U_2 U_3 U_4$)

6b) Bottom Arm ($U_5 U_6 U_7 U_8$)

Fig. 6. Simulation results showing balanced capacitor voltages in M^3C inverter operation

5 Conclusions

This paper presented a new Modular Multilevel Marx Converter, M³C, using modules based on the Marx Generator concept. The addition of one on-off controlled semiconductor switch enabled the parallel connection of capacitors, therefore equalizing their charge.

The M³C concept uses one more controlled semiconductor per cell, but the absence of this extra switch makes the dc voltage balancing possible only in some cases by measuring capacitor voltages and using redundant states, or different cell capacitance values, which makes existing MC cells non-modular.

The M³C cells are modular in design, being suited to build multilevel converters with several tens of levels. They allow a high number of redundant states, which can also be used for capacitor voltage balancing without the need to measure the capacitor voltages or extra balancing algorithms. The drawback of using one extra semiconductor per cell is justifiable by the absence of measurement and control circuits associated with the balancing of capacitor voltages.

To illustrate the M³C operation, as a Marx-generator and as a 5 level inverter, sliding-mode stability based multilevel modulators were applied to a 5 level M³C leg. The sliding-mode stability modulator selected the appropriate levels to synthesize the desired output voltage waveforms. Simulation results showed the needed waveforms and the correct balancing of the dc capacitor voltage waveforms.

References

1. Franquelo, L.G., Rodr guez, J., Leon, J.I., Kouro, S.: The age of multilevel converters arrives. *IEEE Industrial Electronics Magazine* 2(2), 28–39 (2008)
2. Lesnicar, M.R.: An Innovative Modular Multilevel Converter Topology Suitable for a Wide Power Range. In: *IEEE Power Tech. Conference*, Bologna, Italy (2003)
3. Hagiwara, M., Akagi, H.: PWM Control and Experiment of Modular Multilevel Converters. In: *IEEE Power Electronic Specialist Conference*, Rhodes, pp. 154–161 (2008)
4. Adam, G.P., Anaya-Lara, O.G., Burt, M.J.: Comparison between Two VSC-HVDC Transmission Systems Technologies: modular and Neutral Point Clamped Multilevel Converter. In: *35th Annual Conference of the IEEE Industrial Electronics Society – IECON* Porto, Portugal (2009)
5. Adam, G.P., Anaya-Lara, O.G., Finney, S.J., Williams, B.W.: Comparison between flying capacitor and modular multilevel inverters. In: *35th Annual Conference of the IEEE Industrial Electronics Society – IECON* Porto, Portugal (2009)
6. Redondo, L.M., Fernando, S.J.: Repetitive High-Voltage Solid-State Marx Modulator Design for Various Load Conditions. *IEEE Transactions on Plasma Science* 37(8), 1632–1637 (2009)
7. Silva, J., Fernando, P.S.F.: Control Methods for Switching Power Converters. cap. 34. In: Rashid, M.H. (ed.) *Power Electronics Handbook*, 2nd edn., USA, pp. 935–998, p. 1172. Academic Press, Elsevier (2007)
8. Encarna o, L., Silva, J.F.: Sliding Condition Based Sliding Mode Modulators for Multilevel Power Converters. In: *35th Annual Conference of the IEEE Industrial Electronics Society – IECON* Porto, Portugal (2009)
9. Encarna o, L., Silva, J.F.: Reactive Power Compensation Using Sliding-Mode Controlled Three-Phase Multilevel Converters. In: *12th International Conference on Harmonics and Quality of Power – ICHQP* Cascais, Portugal (2006)

Energy Efficient NDMA Multi-packet Detection with Multiple Power Levels

Francisco Ganhão^{1,3}, Miguel Pereira^{1,3}, Luis Bernardo¹, Rui Dinis^{1,3}, Rodolfo Oliveira¹, Paulo Pinto¹, Mário Macedo^{2,4}, and Paulo Pereira^{4,5}

¹ CTS, Uninova, Dep. de Eng. Electrotécnica, Faculdade de Ciências e Tecnologia, FCT, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal

² Dep. de Eng. Electrotécnica, Faculdade de Ciências e Tecnologia, FCT, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal

³ IT, Instituto de Telecomunicações, Portugal

⁴ INESC-ID, Rua Alves Redol, 9. 1000-029 Lisboa, Portugal

⁵ Instituto Superior Técnico, Av. Rovisco Pais. 1049-001 Lisboa, Portugal

fjs.ganhao@gmail.com, miguel.pereira.pro@gmail.com,
{lflb, rdnis, rado, pfp, mmm}@fct.unl.pt, prbp@inesc.pt

Abstract. Multi-packet detection approaches handle packet collisions and errors by forcing packet retransmission and by processing the resulting signals. NDMA (Network Diversity Multiple Access) detection approach forces Q retrasmssions by all stations when Q stations transmit in one collision slot. Previous work assumed that perfect power control is used, where the average reception power is equal for all stations. In this paper we handle the scenario where no power control is used, and multiple power levels are received. We propose a modification to the basic NDMA (Network Diversity Multiple Access) reception mechanism, where some of the stations may not retransmit its packets all the times.

This paper evaluates the EPUP (energy per useful packet) and the goodput for a saturated uplink. Our analytical results are validated by simulation using joint PHY (physical layer) and MAC (Medium Access Control) behavior. We show that by suppressing some retrasmssions we are able to improve the system's EPUP, practically without degrading the network goodput.

Keywords: multi-packet detection, NDMA, energy-per-useful-packet, goodput.

1 Introduction

The conventional approach to cope with collisions in a wireless channel is to discard all packets involved in the collision and to retransmit them. Multi-packet detection mechanisms use the signals associated to multiple collisions to separate the packets involved. In [1], a multi-packet detection technique was proposed, where all users involved in a collision of Q packets retransmit their packets $Q-1$ times. The medium access control (MAC) protocol proposed was named network diversity multiple access (NDMA). To allow packet separation different phase rotations are employed for different packet retrasmssions. An important drawback of the technique of [1] is

that it is only suitable for flat-fading channels. Due to the linear nature of the receivers of [1], the residual interference levels can be high and/or can have significant noise enhancement. In [2] we proposed a frequency-domain multi-packet receiver that allows an efficient packet separation in the presence of successive collisions for NDMA. This receiver is suitable to severely time dispersive channels and does not require uncorrelated channels for different retransmissions. In [3] we extended NDMA to handle the situation where the multi-packet receiver is not able to handle a collision with all active Mobile Terminals (MTs). Yu et al. [4] proposed SICTA, an alternative approach that combines successive interference cancellation (SIC) with a tree algorithm (TA). In SICTA the collided packet signals are used to extract the individual packets, assuming a fixed flat-fading channel. In all cases, it was assumed perfect power control, i.e. equal average power level received at the base station (BS) from all users. However, the power of the received signal from a MT near the BS can be much higher than from the MTs further away of the BS.

This paper studies the goodput and the energy efficiency of NDMA when the BS receives different average power levels at the receiver BS from each user. It analyses the average energy used in the transmission of a successfully received packet for each MT i , i.e. the energy-per-useful-packet of MT i ($EPUP_i$). The paper shows that energy efficiency can be improved if the users whose received signals are more powerful do not retransmit all times, without degrading the saturation throughput and the $EPUP$ for the remaining users. The system overview, including the packet detection and the MAC protocol proposed are presented in sec. 2. The systems' performance are analyzed in sec. 3 and a set of performance results is presented in sec. 4. Finally, sec. 5 is concerned with the conclusions of this paper.

2 Contribution to Sustainability

Energy efficiency is an emerging topic concerning the sustainability of a wireless communication. As a contribution to this topic, the research here presented applies an alternative medium access scheme to NDMA, assuming a non perfect power control, as opposed to other research works. With this scheme, the network's goodput does not degrade, maintaining a good energy efficiency.

3 System Overview

In this paper we consider the saturated uplink transmission in structured wireless systems employing Single Carrier with Frequency-Domain Equalization (SC-FDE) schemes, where a set of MTs send data to a BS. MTs are low resource battery operated devices whereas the BS is a high resource device, which runs the multi-packet detection algorithm in real-time. MTs have a full-duplex radio and employ the NDMA algorithm to send data frames using the time slots defined by the BS (for the sake of simplicity, it is assumed that the packets associated to each user have the same number of bytes, L_{data}). The BS uses the downlink channel to acknowledge transmissions and, possibly, to force packet retransmissions or block the transmission of new packets in the next slot. It is assumed that different data frames arrive

simultaneously and that perfect channel estimation, user detection, and synchronization between local oscillators exists. Data packets are composed of N_{FFT} Fast Fourier Transform (FFT) blocks and have a physical preamble overhead of $N_{PhyPreamble}$ symbols. Each FFT block carries N_{Block} symbols. The physical preamble is used to estimate the channel, synchronize the reception and detect the users involved in a given collision.

3.1 Receiver Structure

To achieve the separation of multiple data frames involved in a collision, we need to have multiple versions of each data frame involved in a collision. Classical NDMA requires Q copies when Q data frames are involved in a collision, but less copies are required when SIC is combined with NDMA. When Q data frames are involved in a collision, the BS forces each MT to retransmit its frame up to $Q-1$ times. Let's assume that the packets are ordered by the reception power, such that $P_q^r \geq P_{q+1}^r$. The MT q retransmits its packet (the q th packet) N_q times, where $q-1 \leq N_q \leq Q-1$. N_q is defined by the BS. Therefore, the receiver has N_q+1 versions of the signal of the q th packet, and jointly detects all frames involved. We consider an iterative receiver that jointly performs the equalization and multipacket detection procedures, where each iteration consists of Q detection stages, one for each frame.

When detecting a given packet we remove the residual interference from the other packets. For the detection of the q th packet and the i th iteration we use Q frequency-domain feedforward filters, each one associated to the signal of a given collision (i.e., one retransmission), and Q frequency-domain feedback filters, each one using the average value of the data signal associated to each packet.

The k th frequency-domain sample associated to the q th packet is $\tilde{A}_{k,q} = \sum_{r=1}^Q F_{k,q}^{(r)} Y_k^{(r)} - \sum_{q'=1}^Q B_{k,q'}^{(q')} \bar{A}_{k,q'}$. The average values $\bar{A}_{k,q'}$ are obtained as follows. The block $\{\bar{A}_{k,q'}; k = 0, 1, \dots, N-1\}$ is the discrete Fourier transform (DFT) of the block $\{\bar{a}_{n,q}; n = 0, 1, \dots, N-1\}$, where $\bar{a}_{n,q}$ denotes the average symbol values conditioned to the FDE output that can be computed as in [5]. It is shown in [2] that the optimal feedforward and feedback coefficients (selected to minimize the ‘signal-to-noise plus interference ratio’, for a given packet and a given iteration) are given by $F_{k,q}^{(r,i)} = \tilde{F}_{k,q}^{(r,i)} / \gamma_q^{(i)}$, with $\gamma_q^{(i)} = \frac{1}{N} \sum_{k=0}^{N-1} \sum_{r=1}^Q \tilde{F}_{k,q}^{(r,i)} H_{k,q}^{(r)}$, and $\tilde{F}_{k,q}^{(r,i)}$ obtained from the set of Q equations:

$$(1 - |\rho_q^{(i)}|^2) H_{k,q}^{(r)*} \sum_{r'=1}^Q \tilde{F}_{k,q}^{(r',i)} \tilde{F}_{k,q}^{(r',i)} + \sum_{q' \neq q} (1 - |\rho_q^{(i)}|^2) H_{k,q'}^{(r)*} \sum_{r'=1}^Q \tilde{F}_{k,q'}^{(r',i)} H_{k,q'}^{(r')} + \alpha \tilde{F}_{k,q}^{(r,i)} = H_{k,q}^{(r)*}, r = 1, 2, \dots, Q \quad (1)$$

The feedback coefficients are then given by $B_{k,q}^{(q',i)} = \sum_{r=1}^Q F_{k,q}^{(r)} H_{k,q'}^{(r,i)} - \delta_{q,q'} (\delta_{q,q'} = 1 \text{ if } q=q' \text{ and } 0 \text{ otherwise})$. Clearly, the Q feedback coefficients are used to remove

interference between packets (as well as residual inter-symbol interference for the packet that is being detected). The feedforward coefficients are selected to minimize the overall noise plus the residual interference due to the fact that we do not have exact data estimates in the feedback loop. For high Signal-to-noise ratios (SNR) and when we do not have any information about the data (i.e., when $\rho_q = 0$), the system of equations (1) gives the $F_{k,q}^{(r,i)}$ coefficients required to orthogonalize the other packets when detecting the q th packet.

The system of equations inherent to (1) might not have a solution or it can be ill conditioned if the correlation between channels associated to different retransmissions is high. If the channel changes significantly from retransmission to retransmission (e.g. if different frequency channels are used for different retransmissions) this correlation can be very low. For systems where this is not practical, we consider the method proposed in [2] where the frequency domain block associated to the r th retransmission of the q th packet, $\{A_{k,q}^{(r)}; k = 0, 1, \dots, N-1\}$, is a cyclic-shifted version of $\{A_{k,q}; k = 0, 1, \dots, N-1\}$, with shift ζ_r . In this paper we assume that the different ζ_r are the odd multiples of $N/2, N/4, N/8, \dots$ [2]. For severely time-dispersive channels this allows a small correlation between different $H_{k,q}^{(r)}$, for each frequency (naturally, as we increase r we increase the correlations). Moreover, envelope fluctuations on the time-domain signal associated to $\{a_{n,q}^{(r)}; n = 0, 1, \dots, N-1\}$ are not too different from the ones associated to $\{a_{n,q}; n = 0, 1, \dots, N-1\}$.

3.2 MAC Protocol

The proposed MAC protocol follows the basic full-duplex NDMA approach [1][2]. The uplink slots are organized as a sequence of collision resolution periods, known as epochs. At the beginning of the epoch, any MT with data frames transmits. No new MTs are allowed to contend until the end of the epoch. An epoch lasts a number of slots equal to the number of MTs transmitting packets. When more than one station transmits, the BS uses the downlink channel to transmit an ACK frame at the end of the first slot, forcing the MTs to retransmit the data frames.

The basic NDMA only requires one bit to force the $Q-1$ retransmissions for every MT. In order to control the number of individual retransmissions, the ACK frame must include the list of MAC addresses detected, and the array with the N_q values, which specify the minimum number of retransmissions required for each colliding MT. Failed data frames are retransmitted in the next epochs up to M_R times before being dropped. In the following analysis we do not take into account the duration of the ACK frame, which is transmitted between data frames.

4 Performance Analysis

4.1 Goodput Analysis

The MultiPacket Reception (MPR) system can be characterized by the probability of a MT l successfully transmitting a packet, given by $q_{Q,l}$, $0 \leq q_{Q,l} \leq 1$, with l , $1 \leq l \leq Q$, and

where Q defines the number of MTs colliding. This probability is system specific, and depends on the successful reception of every bit, $ber_{Q,l}$, which is influenced by the set of received powers at the BS and the N_q values. Assuming a fixed size of L_{data} bits per packet, then $q_{Q,l}$ is given by

$$q_{Q,l} = (1 - ber_{Q,l})^{L_{data}}. \quad (2)$$

Let S_l^{sat} be the saturation goodput for MT l . It can be calculated by the ratio of the expected number of packets successfully transmitted by MT l (with up to $R-1$ retransmissions) to the expected packet transmission duration, and is given by

$$\begin{aligned} S_l^{sat} &= \frac{\sum_{r=1}^R q_{Q,l} (1 - q_{Q,l})^{r-1}}{QR(1 - q_{Q,l})^R + \sum_{r=1}^R Qr q_{Q,l} (1 - q_{Q,l})^{r-1}} \\ &= \frac{q_{Q,l} (1 - (1 - q_{Q,l})^R)}{q_{Q,l} QR(1 - q_{Q,l})^R + Q(1 - (1 - q_{Q,l})^R)^2}. \end{aligned} \quad (3)$$

The total channel's saturation goodput is,

$$S^{sat} = \sum_{l=1}^Q S_l^{sat}. \quad (4)$$

4.2 Energy Analysis

For a low power transmission system, the energy consumption model has to consider both the transmission energy and the circuit energy consumption. Cui et al. [6] proposed an energy model for uncoded M-QAM (M-ary quadrature amplitude modulation) that can be adapted to our scenario, considering both, the transmission and reception energy consumption for full-duplex, during the slot time, T_{slot} . Cui et al. show that the energy consumption per packet transmission in MT l is approximately given by $E_l^P \approx (1 + \beta) P_l^t T_{slot} + P_c T_{slot}$, where P_c denotes the total circuit energy consumption and $\beta = P_{amp}/P_l^t$ is the ratio of power consumption on the power amplifier to the transmitted power (P_l^t). This ratio is given by $\beta = \xi/\eta - 1$ with η as the drain efficiency of the radio frequency power amplifier and ξ as the peak-to-average-ratio.

An uncoded scenario with one bit per symbol was considered for this paper due to its extreme simplicity. Thus, $\xi = \frac{\max(b_n b_n^*)}{E[b_n b_n^*]} = 1$. Assuming the K th-power path-loss model at distance d (meters), the transmission power is expressed as $P_t = P_r G_1 d^K M_l$, where P_r is the received power, M_l is the link margin compensating

the hardware process variations and other additive noise, and G_1 is the gain factor at $d=1$ m. An AWGN power spectral density of $\sigma^2 = -174$ dBm/Hz was considered for a given bandwidth B .

Assuming $P_r = ME_b/T_{slot}$, the expended energy for each packet is defined as

$$E_l^p \approx (1 + \beta)G_1 d^\kappa M_l ME_b + P_c T_{on}. \quad (5)$$

The energy per useful packet of MT l , denoted $EPUP_l$, measures the average transmitted energy for a correctly received packet. It is given by:

$$\begin{aligned} EPUP_l &= \frac{E_l^p R (1 - q_{Q,l})^R + \sum_{r=1}^R E_l^p r q_{Q,l} (1 - q_{Q,l})^{r-1}}{1 - (1 - q_{Q,l})^R} \\ &= E_l^p \left(\frac{R (1 - q_{Q,l})^R}{1 - (1 - q_{Q,l})^R} + \frac{1 - (1 - q_{Q,l})^R}{q_{Q,l}} \right). \end{aligned} \quad (6)$$

Therefore, the channel's average $EPUP$, is

$$EPUP = \left(\sum_{l=1}^Q EPUP_l \right) / Q. \quad (7)$$

5 Simulations

The current section presents a performance analysis of the proposed system model for a flat channel, assuming a saturated system. The results were measured using Monte Carlo simulations. Table 1 presents the energy model parameters.

The results hereof, assume two groups of nodes: A with Q_1 MTs and B with Q_2 . The former group transmits with maximum transmission power, while the latter transmits with 40%. A base station receives packets from both groups and manages the scheduling of collided packets. A packet collision, might include nodes from both groups or just one. The receiver's structure supports up to $Q=4$ collisions. For simplicity purposes, it is assumed that at most one node from group A, transmits at the same time with up to $Q-1$ nodes from group B. Nodes are spatially distributed around the base station at a distance of 20m.

Table 1. Simulation parameters

Parameter	Value	Parameter	Value	Parameter	Value
P_c	164.998 mW	ξ	1	η	0.35
T_{slot}	0.0205 ms	B	2.5 MHz	L_{data}	1024 bit
G_1	30 dB	κ	3.5	M_l	40 dB

Figure 1, presents the bit error rate (BER) of an AWGN channel. It shows that the BER improves for an increasing number of collisions (higher Q). The BER reduction is more significant when the interfering signals have lower power. Therefore, the proposed iterative receiver is capable of removing interference with multiple receiver power levels.

Figure 2 illustrates the energy per useful packet (EPUP). For a fixed number of Q=4 collisions, it is observable that an increasing number of retransmissions for each epoch, degrade the EPUP for a saturated channel, the same applies for nodes of a weaker transmission power. Correlating the EPUP with the goodput from Figure 3, it is conclusive that the optimum, or minimum, EPUP is obtained once the goodput reaches its maximum. The goodput is also affected for nodes with a weaker transmission power.

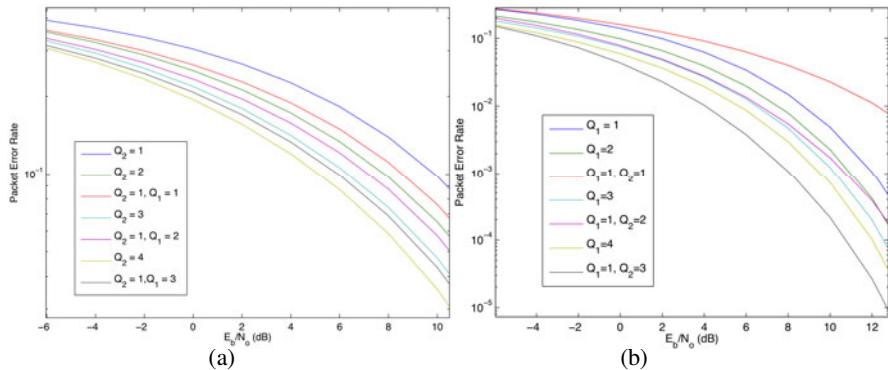


Fig. 1. Bit error rate: (a) Q_2 nodes that transmit with 40% of the maximum transmission power; (2) Q_1 nodes that transmit with 100% of the maximum transmission power

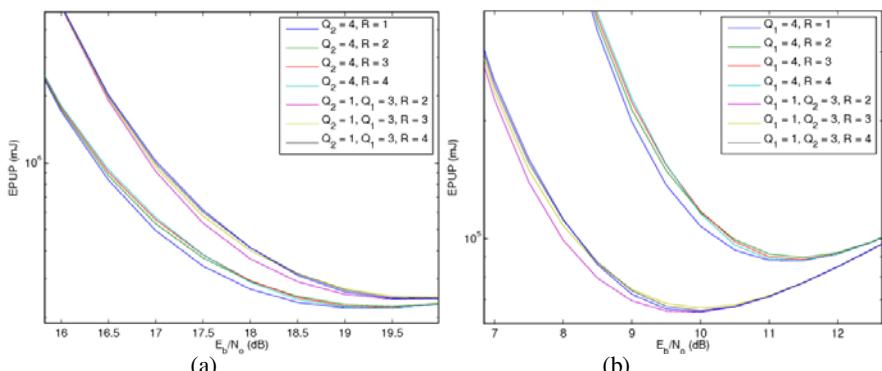


Fig. 2. Energy per useful packet: (a) Q_2 nodes that transmit with 40% of the maximum transmission power; (2) Q_1 nodes that transmit with 100% of the maximum transmission power

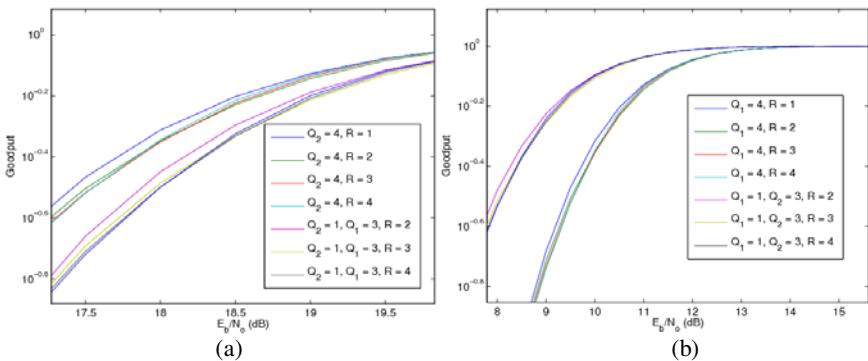


Fig. 3. Goodput: (a) Q_2 nodes that transmit with 40% of the maximum transmission power; (2) Q_1 nodes that transmit with 100% of the maximum transmission power

6 Conclusions

In this paper, we have proposed a new approach to handle differentiated power levels at an MPR receiver and analysed the resulting energy efficiency and goodput. The paper shows that the suppression of retransmissions of the most powerful signals increases the system goodput and reduces the EPUP, showing that this approach allows a performance improvement of NDMA over classic NDMA with perfect power control. For future research, it is intended to enhance the analytical model for an unsaturated channel and study its performance with upper layer protocols. Other research options are also available, such as enhancing the proposed architecture in terms of scheduling and the receiver's structure.

References

1. Tsatsanis, M.K., Ruifeng, Z., Banerjee, S.: Network-assisted diversity for random access wireless networks. *IEEE Transactions on Signal Processing* 48(3), 702–711 (2000)
2. Dinis, R., Montezuma, P., Bernardo, L., Oliveira, R., Pereira, M., Pinto, P.: Frequency-domain multipacket detection: a high throughput technique for SC-FDE systems. *IEEE Trans. on Wireless Communications* 8(7), 3798–3807 (2009)
3. Pereira, M., Bernardo, L., Dinis, R., Oliveira, R., Carvalho, P., Pinto, P.: A MAC Protocol for Half-Duplex Multi-Packet Detection in SC-FDE Systems. In: *IEEE Vehicular Technology Conference (VTC)-Spring*, IEEE Press, Barcelona (2009)
4. Yu, Y., Giannakis, G.B.: SICTA: a 0.693 contention tree algorithm using successive interference cancellation. *IEEE/ACM Trans. on Networking* 3, 1908–1916 (2005)
5. Dinis, R., Carvalho, P., Martins, J.: Soft Combining ARQ Techniques for Wireless Systems Employing SC-FDE Schemes. In: *17th Int. Conf. on Computer Communications and Networks*, pp. 1–5. IEEE Press, St. Thomas U.S. Virgin Islands (2008)
6. Cui, S., Goldsmith, A.J., Bahai, A.: Energy-constrained modulation optimization. *IEEE Trans. on Wireless Communications* 4(5), 2349–2360 (2005)
7. NS-2 Network Simulator, version 2.34 (2009), <http://www.isi.edu/nsnam/ns/>

Resistive Random Access Memories (RRAMs) Based on Metal Nanoparticles

Asal Kiazaedeh, Paulo R. Rocha, Qian Chen, and Henrique L. Gomes

Center of Electronic, Optoelectronic and Telecommunications, Faculdade de Ciências e Tecnologia, Universidade do Algarve, Campus de Gambelas,
8005-139 Faro, Portugal

Abstract. It is demonstrated that planar structures based on silver nanoparticles hosted in a polymer matrix show reliable and reproducible switching properties attractive for non-volatile memory applications. These systems can be programmed between a low conductance (off-state) and high conductance (on-state) with an on/off ratio of 3 orders of magnitude, large retention times and good cycle endurance. The planar structure design offers a series of advantages discussed in this contribution, which make it an ideal tool to elucidate the resistive switching phenomena.

Keywords: Nanoparticles, resistive switching, non-volatile memory.

1 Introduction

Non-volatile memories (NVMs) have become a major technology in the storing of digital information. Flash memory is currently the main stream of the non-volatile memory technology and can be found everywhere in our daily life particularly in portable devices. The absence of mechanical parts, lighter weight and lower power dissipation makes flash NVMs ideal for these applications. In flash memories the information is stored in the form of charge contained in a so-called floating gate (FG) completely surrounded by a dielectric (hence the name and located between the channel region and the conventional, externally accessible, gate of a field effect device (FET)). The amount of charge stored on the FG layer can be easily sensed since it is directly proportional to the threshold voltage of the FET. Flash retention relies on charge-storing on a floating gate, and this lays a challenge for the scaling-down. For instance, less than 100 electrons will be stored in 32 nm node, the information will be easily lost due to the leakage through the thin dielectric. Furthermore, as CMOS scaling proceeds, there is an increasing need to simplify and unify different technologies. In fact, to improve system performance, circuit designers often combine several memory types, adding complexity and cost. The unification of these different technologies would allow further scaling and the decrease of the system cost. Many of the memory caches in the hierarchy of today's computer architecture could be eliminated, reducing cost and complexity. Ideally, the ultimate universal memory aims to replace conventional embedded and stand-alone memories. Alternative routes to traditional memories are thus under intense investigation, with new NVM concepts and storage principles being investigated that may overcome the intrinsic limitation of flash and allow unification of existing memories. A variety of next generation NVMs has been proposed, from which Resistive Random Access

Memory (RRAMs) are the latest. RRAMs are being intensively investigated by companies and universities worldwide. The main advantages offered by the technology are higher packing density, faster switching speed and longer storage life. The appeal of RRAM is that each element is a two-terminal device. Switching between high and low resistance is achieved by means of an appropriate electrical pulse, and the read-out process consists of applying a (lower) voltage to probe the state of the resistance. This type of element can be incorporated into cross-point arrays. Resistive type memories are usually metal-insulator-metal (MIM) structures which show non-volatile electrically induced resistance variations of up to nine orders of magnitude. Insulating materials can be as diversified as CuO, CoO, ZnO, NiO, TiO₂, MgO, Al₂O₃, SiO₂, perovskites, among others [1-6]. Thin films incorporating metallic nanoparticles (NPs) also show reliable and reproducible switching properties. Nanoparticles can be capped into polymeric materials and stable solutions are readily available [7-9]. Thin films hosting a well-defined distribution of nanoparticles are easily fabricated by spin-coating, printing, or dip-coating techniques offering the prospect of low fabrication cost, mechanical flexibility and lightweight.

2 Technological Innovation for Sustainability

Together with development of the microelectronic industry, the research on low power, low cost, and high density memory devices is necessary for the digital systems, especially for the portable systems. For instance, the absence of mechanical parts, lighter weight and lower power dissipation makes flash non-volatile memory ideal for these applications. However, flash retention relies on charge-storing on a floating gate, and this lays a challenge for the scaling-down. In this work we demonstrate a memory device, which makes use of thin polymer film hosting a matrix of silver nanoparticles. With on/off ratio as high as 10³, a large retention time and good cycle endurance, the nanoparticles based device is a serious candidate to replace currently available non-volatile memories, and is able to improve all the relevant components as a reliable memory device.

Planar structures have a significant lower intrinsic capacitance than sandwich-type structures commonly reported; therefore, they should have a faster dynamic response. In addition, these co-planar structures have the active layer directly exposed and can be probed by a number of surface analytical techniques to identify and characterize topographical, morphological changes that occur in the devices upon resistive switching.

3 System Description

The colloidal solution of PolyVinylPyrrolidone(PVP) capped- Silver nanoparticles was deposited on top of preformed gold microelectrode arrays fabricated on thermal oxidized silicon wafers. The interdigitated gold microelectrode arrays were separated apart 10 μm and 10.000 μm long. The samples were dried at room temperature for several hours for removal of the solvent. Prior to electrical measurements the samples were also pumped in vaccum to remove further any residual solvent. Electrical measurements were carried out using a Keithley 487 picoammeter/voltage source in dark conditions, high vacuum. Figure 1 shows the device test structure used to measure the electrical properties of the system comprised of nanocrystals in a semi

insulating or insulating host matrix. A thin film is formed between the two gold electrodes on top of the insulating silicon dioxide surface by drop casting or by spin coating. During all the measurement the conductive silicon substrate is kept grounded to prevent charging of the SiO_2 layer.

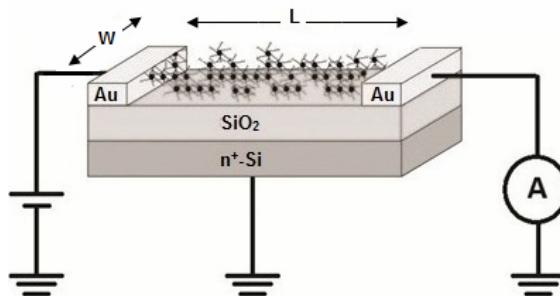


Fig. 1. Planar device structure with two gold electrodes. The device dimensions are $W=10.000\mu\text{m}$ and $L=10\mu\text{m}$.

3.1 Current-Voltage Characteristics

Before the devices show resistive switching behavior, they have to undergo a forming process that is induced by applying a high bias voltage. In practice, forming voltages are typically $\sim 50\text{V}$. After forming the device exhibits the bistable current-voltage characteristics represented in Fig. 2. A low conductance state named off-state and a high conductance state designated as on-state. The on-state has a symmetric negative differentially resistance (NDR) region located between 25 and 30 V. The memory can be switched between off and on-state by applying voltage pulses with amplitudes corresponding to the top and bottom of the NDR region, in Fig. 2 about 20 V and 40 V, respectively.

The device can be read out non-destructively using voltages $< 10\text{V}$. The programming voltages are larger than those observed in conventional sandwich capacitor-like devices because the distance between electrodes is also significant higher.

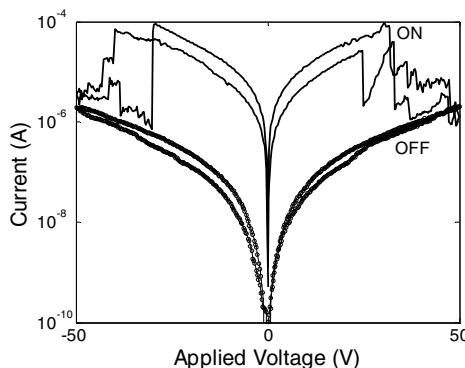


Fig. 2. Current-voltage characteristics showing the high conductance on-state and low conductance off-state

3.2 The Memory Characteristics

The nanoparticle based memory is nonvolatile, retaining the programmed conductance state for at least a week without any applied bias as long as they are stored in vacuum or in inert atmosphere. Fig.3 (a) shows the retention time for both states using a read voltage of 3 V.

Once the device is in a given memory state, it can be reliably read many times without degradation. Shown in Fig. 3(b) is a segment of the current response to write-read-erase-read voltage waveforms. The write voltage (W) is 35 V and the erase (E) is 50 V. For both states the read voltage (R) is 3V.

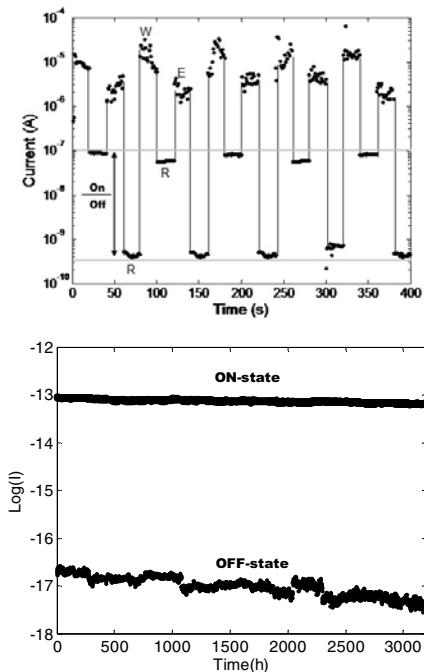


Fig. 3. (a) retention time of two memory states has been obtained at 3 volt. (b) typical current response to the write-read-rewrite-read voltage cycles. The write voltage (W) is 35 V and the erase (E) is 50 V. For both states the read voltage (R) is 3V.

4 Conclusion

Metal nanoparticles embedded in a polymer matrix exhibit the ability to switch between different nonvolatile conductance states. Programming of the memory states is done using voltage pulses. The memory device concept was demonstrated using a lateral structure with a large gap between electrodes ($10\mu\text{m}$). This structure needs substantially high programming voltages. However, it is feasible to bring the device dimensions to a nanometer scale and bring the magnitude of the operating voltages to values compatible with CMOS technology.

Since the basic active layer is a soluble polymer, large area arrays of memories can easily be fabricated into flexible substrates, thus opening the prospect for low-cost production of memory arrays for instance for radio-identification tags.

Acknowledgement. We gratefully acknowledge the financial support received from the Dutch Polymer Institute (DPI), project n.^o 703, from Fundação para Ciência e Tecnologia (FCT) through the research Unit, Center of Electronics Optoelectronics and Telecommunications (CEOT), REEQ/601/EEI/2005 and the POCI 2010, FEDER and the organic chemistry laboratories in Algarve University.

References

1. Hickmott, T.W.: Low-Frequency Negative Resistance in Thin Anodic Oxide Films. *Journal of Applied Physics* 33(9), 2669–2682 (1962)
2. Hickmott, T.W.: Electron Emission, Electroluminescence, and Voltage-Controlled Negative Resistance in Al-Al₂O₃-Au Diodes. *Journal of Applied Physics* 36(6), 1885–1896 (1965)
3. Barriac, C.: Study of the electrical properties of Al-Al₂O₃-metal structures. *Physica Status Solidi A - Applied Research* 34(2), 621–633 (1962)
4. Simmons, J.G., Verderber, R.R.: New Conduction and Reversible Memory Phenomena in Thin Insulating Films. *Proceedings of the Royal Society of London. Series A* 301, 77–102 (1967)
5. Sutherland, R.R.: Theory for Negative Resistance and Memory Effects in Thin Insulating Films and Its Application to Au-ZnS-Au Devices. *Journal of Physics D - Applied Physics* 4(3), 468–479 (1971)
6. Ansari, A.A., Qadeer, A.: Memory Switching in Thermally Grown Titanium-Oxide Films. *Journal of Physics D –Applied Physics* 18(5), 911–917 (1985)
7. Ma, L.P., Liu, J., Pyo, S., Xu, Q.F., Yang, Y.: Nonvolatile electrical bistability of organic/metal-nanocluster/organic system. *Applied Physics Letters* 82(9), 1419–1421 (2003)
8. Paul, S., Pearson, C., Molloy, A., Cousins, M.A., Green, M., Koliopoulos, S., Dimitrakis, P., Normand, P., Tsoukalas, D., Petty, M.C.: Langmuir-Blodgett film deposition of metallic nanoparticles and their application to electronic memory structures. *Nano Letters* 3(4), 533–536 (2003)
9. Silva, H., Gomes, H.L., Pogorelov, Y.G., Stallinga, P., de Leeuw, D.M., Araujo, J.P., Sousa, J.B., Meskers, S.C.J., Kakazei, G., Cardoso, S., Freitas, P.P.: Resistive switching in nanostructured thin films. *Applied Physics Letters* 94, 1–3 (2009)

Design, Synthesis, Characterization and Use of Random Conjugated Copolymers for Optoelectronic Applications

Anna Calabrese^{1,2}, Andrea Pellegrino², Riccardo Po², Nicola Perin²,
Alessandra Tacca², Luca Longo², Nadia Camaioni³, Francesca Tinti³,
Siraye E. Debebe^{3,4}, Salvatore Patanè⁵, and Alfio Consoli⁶

¹ Scuola Superiore di Catania, Università di Catania, Via San Nullo 5/I, 95123 Catania, Italy
anna.calabrese@ssc.unict.it

² Research Center for non Conventional Energies Istituto Eni Donegani, ENI S.p.A, Via G.
Fauser 4, 28100 Novara, Italy
{andrea.pellegrino,riccardo.po,nicola.perin,alessandra.tacca,luca.longo}@eni.com

³ Istituto per la Sintesi Organica e la Fotoreattività (CNR-ISOF), Consiglio Nazionale delle
Ricerche, Via P. Gobetti 101, 40129 Bologna, Italy
{camaioni,ftinti,siraye}@isof.cnr.it

⁴ Department of Chemistry, Addis Ababa University, PO Box 1176, Addis Ababa, Ethiopia

⁵ Dipartimento di Fisica e Ingegneria Elettronica, Università di Messina, Via Salita Sperone
31, 98166 Messina, Italy
patanes@unime.it

⁶ Dipartimento di Ingegneria Elettrica Elettronica e dei Sistemi, Università di Catania,
Viale A. Doria 6, 95125 Catania, Italy
aconsoli@diees.unict.it

Abstract. We report the synthesis and the optoelectronic characterization of a new family of random conjugated copolymers based on 9, 9-bisalkylfluorene, thiophene and benzothiadiazole monomers unit synthesized by a palladium-catalyzed Suzuki cross-coupling reaction. The photophysical, thermal, electrochemical properties were investigated. The electronic structures of the copolymers were simulated via quantum chemical calculations. Bulk heterojunction solar cells based on these copolymers blended with fullerene, exhibited power conversion efficiency as high as 1% under illumination of 97 mWcm⁻². One of the synthesized copolymers has been successfully tested as active layer in simple light-emitting diode, working in the green spectral region and exhibiting promising optical and electrical properties. This study suggests that these random copolymers are versatile and are promising in a wide range of optoelectronic devices.

Keywords: conjugated polymer, random, organic solar cell, organic light-emitting diode.

1 Introduction

The field of organic optoelectronics is growing up since the late 1980s [1]. The research has been largely driven by design and development of different types of

functional materials for optoelectronic applications such as organic light-emitting diodes (OLEDs) [2], organic solar cells (OSCs) [3], organic field-effect transistors (OFETs) [4], etc. The common feature of these applications consists of using materials with suitable transport and optical properties. Conjugated polymers are a class of materials that are gaining great attention by the scientific community due to its potential of providing environmentally safe, flexible, lightweight and inexpensive electronics. They show flexibility in synthesis, high yield of charge generation when mixed with electron acceptor materials and good stability. Furthermore, they have relatively high absorption coefficients [5] leading to high optical densities in thin solid films. Among the conjugated polymers, alternating fluorene copolymers (APFO) [6] and their derivatives have been widely used for efficient organic devices. In particular, the APFO-3 polymer has demonstrated efficiencies as high as 4.2%. In this polymer the fluorene unit has been used in conjunction with alternating electron-withdrawing (A) and electron-donating (D) groups. The structure is a strictly *alternating* sequence of fluorene and donor-acceptor-donor (DAD) units. Conjugated *random* copolymers are, by far, less used than alternating copolymers for devices fabrication. While the disordered structure may hamper the crystallization and decrease the carrier mobility, the presence of different monomer units sequences generate a distribution of energy gaps and increase the light harvesting ability. Furthermore an advantage of random copolymers lies in their simplicity of synthesis that consists in a one-pot polymerization step from readily available precursors. In this work we report the synthesis and characterization of a novel family of conjugated copolymers based on same the monomeric units of APFO-3 (F: fluorene, T: thiophene, B: benzothiadiazole) but having a pseudorandom structure. We applied these new copolymers in solar cells and bright green OLEDs.

2 Technological Innovation for Sustainability

The development of sustainable energy production is a topic of growing importance in view of the limited supply of fossil fuels, which is expensive financially and not environmentally friendly. One of the possible sustainable energy sources is the sun, which makes the development of photovoltaic devices interesting. Currently, the market of photovoltaic is dominated by silicon cells technology. However, it is still not cheap enough to allow a wide diffusion in the absence of government incentives. For this reason huge efforts of research and development have been spent to find alternative and improved solutions. Organic devices are lightweight and can be made flexible, opening the possibility for a wide range of applications. Our contribution in this field is the development of promising organic materials for photovoltaic and optoelectronic applications. In particular, our work revolves around the conjugated polymers that have received considerable attention because of their promising performance. In this paper we report the synthesis and characterization of a family of novel conjugated copolymers having a pseudorandom structure. The synthesis of random instead of alternating copolymers has the aim to improve the harvesting of the sunlight. Indeed, the mixture of possible copolymer resulting from the statistical combination (in condition of equal affinity between the monomers that form copolymers) results in a broadening of the absorption bands due to overlapping of different energy levels. This is one of fundamental prerequisites in order to obtain high efficiency photovoltaic devices.

3 Experimental

Three different pseudo-random copolymers (**PFB-co-FT**, **PBT-co-BF**, **PTF-co-TB**) were synthesized by a palladium-catalyzed Suzuki cross-coupling reaction [7] from dibromides and boronic diacids or diesters. More details can be found elsewhere [8]. The weight-average molecular weight (M_w) and polydispersity index (PDI) were measured by gel permeation chromatography (GPC) using THF as eluent and monodisperse polystyrenes as internal standards. UV-Visible absorption spectra of all copolymers was recorded at room temperature with a Lambda 950 spectrophotometer. Electrochemical measurements were performed with an Autolab PGSTAT30 potentiostat/galvanostat in a one compartment three-electrode cell working in argon-purged acetonitrile solution with 0.1 M Bu_4NBF_4 as supporting electrolyte. We used a Pt counter electrode, an aqueous saturated calomel (SCE) reference electrode and a Glassy Carbon working electrode which was calibrated against the Fc^+/Fc (ferricenium/ferrocene) redox couple, according to IUPAC [9]. The films formed on the electrode were analyzed at a scan rate of 200 mV/s. OSCs were fabricated by first spin-coating poly(ethylenedioxothiophene:polystyrenesulfonic acid) (PEDOT-PSS) on top of cleaned, pre-patterned indium-tin-oxide (ITO) coated glass substrates as polymer anode (50 nm in thick). The anode polymer film was then annealed at 120°C for 10 min. The copolymers blended with PCBM in solution were deposited on the top of PEDOT:PSS by spin-coating. The cathode consisting of Al (70 nm) was then sequentially deposited on top of the active layer by thermal evaporation in vacuum lower than 10^{-6} Torr giving a sandwich solar cell structure of ITO/PEDOT:PSS/Copolymer:PCBM/Al. Current density-voltage characteristics were measured using a Solar Simulator Abet 2000 (Class A, AM1.5G). All characterization was carried out in an ambient environment. OLEDs were fabricated by the same procedure as that of solar cells but in this case the only copolymer in solution were deposited on the top of PEDOT:PSS by spin-coating. The corresponding configuration of devices was ITO/PEDOT:PSS/Copolymer/Al. Steady-state Photoluminescence (PL) spectra were recorded at room temperature with an Fluorolog 3 spectrofluorometer. The steady state current-voltage (I-V) characteristic was recorded by NI-USB 6229 National Instruments acquisition card. The electroluminescence (EL) spectra were collected installing the device inside the sample chamber of an IF 650 (Pelkin Elmer) spectrophotometer, working in emission mode, by injecting a current of 1 mA. To this purpose a HP E3631 triple output power supply was used as a constant current generator.

4 Modeling

In order to understand the correlations between property and structure, theoretical calculations were carried out. The calculations of HOMO/LUMO distribution of copolymers were performed in two steps. The monomeric units was first modeled by using a quantum mechanical Hartree Fock methods [10]. In the second step we obtained the geometric optimization of monomers and oligomers by a density functional theory approach [11]. The performed calculations showed how HOMO and LUMO positions are affected by electron-donating and electron-withdrawing groups. The last occupied orbital is localized primarily on the electron-rich units of the

molecule (Fluorene and Thiophene) while the first empty orbital is essentially localized on the electron-poor unit (Benzothiadiazole). This means that the molecular orbitals are not extended to the whole structure, but are confined in a limited region of space. This has important implications both in the exciton formation and in the charge transfer. Following the model results we can conclude that the charge *hopping* among localized sites is one of the main transport mechanism. In order to obtain an electronic transfer between the copolymer and the PCBM, the latter should be close to a benzothiadiazole unit. On the other hand, the fact that a partial charge separation already exists, may promote the formation of excitons and extend their lifetime.

5 Results and Discussion

Material synthesis: The chemical structures of all synthesized copolymers is showed in Fig. 1. In the three copolymers, in turn, each F, T, B comonomer unit is alternated to the other two units, which are randomly distributed. The weight-average molecular weight (Mw), and polydispersity indices (PDIs) are summarized in Table 1.

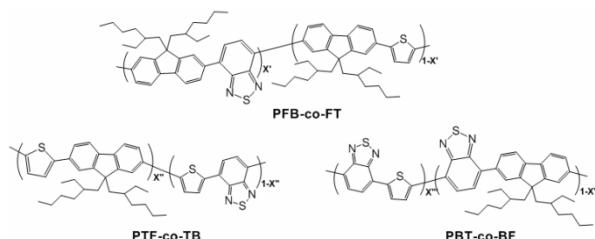


Fig. 1. Chemical structure of the synthesized copolymers

Table 1. Molecular weights, optical and electrochemical data of copolymers

Entry	GPC		UV-Vis absorption		Cyclic Voltammetry		
	Mw (g/mol)	PDI	Eg ^{opt} (eV)	λ _{max} (nm)	E _{on} ^{ox} /HOMO (V)/(eV)	E _{on} ^{red} /LUMO (V)/(eV)	Eg ^{cv} (eV)
PFB-co-FT	34600	3.00	2.38	321/460	0.81/-5.61	-1.93/-2.87	2.74
PTF-co-TB	3200	1.80	1.86	402/552	0.51/-5.31	-1.54/-3.26	2.05
PBT-co-BF	1600	2.60	1.40	326/526	0.21/-5.01	-1.40/-3.40	1.61

The properties of conjugated polymers are remarkably sensitive to the presence of impurities, which might act as uncontrolled dopants, traps of charge carriers, quenchers of excited states, etc. To improve the performance the copolymers solution was treated by ammonia and ethylenediaminetetraacetic acid to remove contaminants such as catalyst residue and side-products from copolymers synthesis. (See Table 2).

Optical properties: Absorption spectroscopy provides information about the spectral coverage and the magnitude of the optical energy gap that are an important

parameters in device designing. The UV-Vis absorption spectra of the copolymers in thin films are shown in Fig. 2.

Table 2. Elemental analysis of copolymers at different stages of purification

Entry	Before purification		After purification	
	Pd (ppm)	P (ppm)	Pd (ppm)	P (ppm)
PFB-co-FT	3400	800	10	700
PTF-co-TB	1100	800	100	800
PBT-co-BF	30000	9000	5000	900

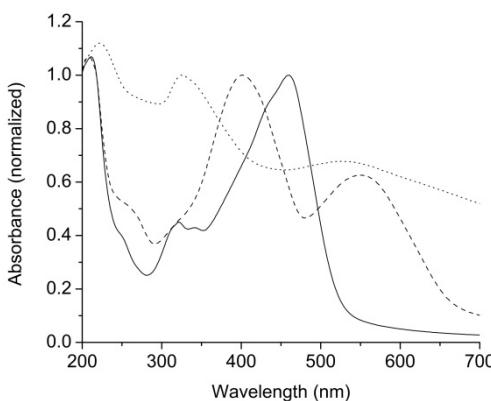


Fig. 2. Absorption spectra of films of random copolymers **PFB-co-FT** (solid curve), **PTF-co-TB** (dash curve) and **PBT-co-BF** (dot curve), normalized at the band around 400 nm

The optical energy gap of the copolymers was estimated from the onset of absorption and data are reported in Table 1.

Electrochemical properties: Cyclic voltammetry (CV) was employed to estimate the HOMO and LUMO energy levels of copolymers. Electrochemical data were calculated from the onsets of oxidation and reduction potentials [12]. The data obtained are reported in Table 1 and schematized in Fig. 3.

The electrochemical gap is greater than the optical gap calculated from the UV-Vis spectra. This discrepancy is usually related to the charge carriers formation in voltammetric measurements [13]. All the copolymers have HOMO and LUMO levels higher than the commonly used [6,6]-phenyl-C₆₁-butyric acid methyl ester (PCBM) acceptor [14] (see Fig. 3). **PFB-co-FT** exhibits the higher molecular weight and the higher energy gap compared to the other two copolymers. Furthermore is the copolymer with the highest E_{HOMO}donor-E_{LUMO}acceptor difference. On this basis **PFB-co-FT** is expected to lead to photovoltaic devices with the better V_{oc} [15]. The E_{HOMO}donor-E_{LUMO}acceptor difference of **PTF-co-TB** is lower, but the smaller energy gap could compensate the expected lower V_{oc} with a higher J_{sc} in designing an efficient solar cell. **PBT-co-BF**, is the copolymer with the lower energy gap and this would make it the best candidate for OSCs.

5.1 Conjugated Polymer-Based Photovoltaic Devices

Photovoltaic devices were fabricated in a typical sandwich structure of glass-ITO/PEDOT:PSS/Copolymer:PCBM/Al, using copolymers as electron donors and the PCBM as electron acceptor. Photovoltaic characterization includes different approaches to optimize the efficiency of the devices. The parameters taken into account were: selection of the best solvent, optimization of D:A weight ratio, optimization of the active layer thickness and analyzing annealing effects. We have investigated the effects of three different solvents on the photovoltaic performance: Chloroform (CF), Chlorobenzene (CB) and orto-DiChloroBenzene (o-DCB), individually and in a mixture form. The copolymer **PBT-co-BF** showed very low solubility and the tendency to agglomerate in all solvents; this made difficult to obtain a good film and accordingly devices. In order to choose the best blend composition, active layers with D:A w/w ratios 1:1, 1:2, 1:3, and 1:4 were studied. The photovoltaic parameters of the best cells are summarized in Table 3.

Table 3. Photovoltaic parameters of the best solar cells. The active area was 0.22cm^2 .

Donor	Acceptor	D/A ratio (w/w)	Solvent	C (mg/ml)	Voc (V)	Jsc (mAcm $^{-2}$)	FF	η (%)
PFB-co-FT	PCBM	1:2	CB:CF (1:1 v/v)	10	0.92	0.90	0.40	0.35
PTF-co-TB	PCBM	1:4	CB	5	0.94	3.43	0.31	1.08

Fig. 3 shows the J-V curve of the best devices in the dark and under AM 1.5G simulated sunlight at an intensity of 97 mWcm^{-2} . The active layer thickness was 100 nm for all cells. The effects of thermal annealing were also studied but it has shown a negative effect in all cases.

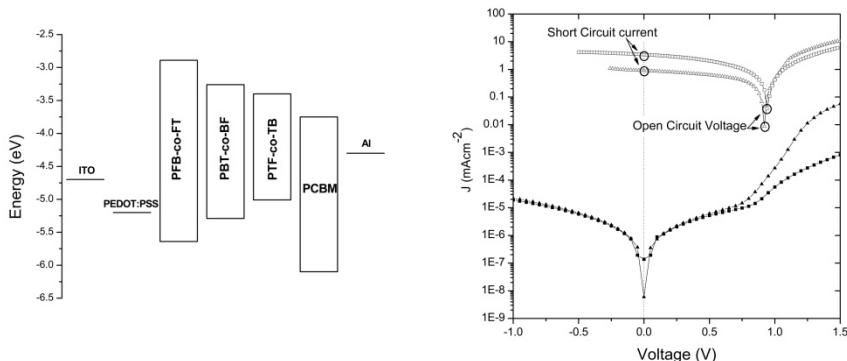


Fig. 3. Energy level diagram of the device components (Left); J-V curve of the best solar cells for **PFB-co-FT** (triangle) and **PTF-co-TB** (square) in the dark (black) and under illumination (white) (Right)

5.2 Light-Emitting Diode Based on Conjugated Polymer

Among all copolymers investigated, **PFB-co-FT** has been recognized as promising material for green emitters due to its photoluminescence efficiency. The normalized UV-Vis and PL spectra of copolymer are shown in Fig. 4; the absorption maxima is at 460 nm while the emission maxima is at 543 nm. The devices was fabricated in a simple structure of glass/ITO/PEDOT:PSS/Copolymer/Al. The EL spectra of the device, is showed in Fig. 4.

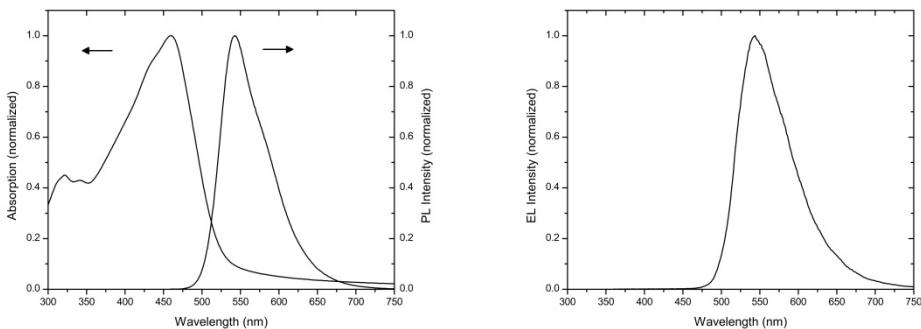


Fig. 4. Normalized UV-Vis absorption and PL spectra for the **PFB-co-FT** copolymer in film (*Left*); Normalized EL spectra of the device recorded at 7.65 V and 1mA (*Right*)

After one month of storage at room temperature the device was re-examined, with identical results and without degradation. This outcome confirms the good stability of the film.

6 Conclusions

In this work we have synthesized a novel family of random copolymers. The optical and electrochemical behaviors of the materials were measured and the optoelectronic performance were tested. The results clearly demonstrated that there is a correlation between the chemical structure and the HOMO levels determined from electrochemical studies as well as the V_{oc} values determined from photovoltaic devices. **PTF-co-TB** showed interesting photovoltaic properties with V_{oc} of 0.94 V and an efficiency over 1%. The **PFB-co-FT** copolymer was tested as active layer in a simple OLED. The device emits in the green and might be a promising candidate for electroluminescent materials due to its excellent luminescent properties, solubility, film-forming property and stability. We can conclude that this class of copolymer materials is very promising for optoelectronic applications.

References

1. Tang, C.W.: Two-layer organic photovoltaic cell. *Appl. Phys. Lett.* 48, 183–185 (1986)
2. Burroughes, J.H., Bradley, D.D.C., Brown, A.R., Marks, R.N., Mackay, K., Friend, R.H., Burns, P.L., Holmes, A.B.: Light-emitting diodes based on conjugated polymers. *Nature* 347, 539–541 (1990)

3. Wöhrle, D., Meissner, D.: *Organic Solar Cells*. *Adv. Mater.* 3, 129–138 (1991)
4. Drury, C.J., Mutsaers, C.M.J., Hart, C.M., Matters, M., Leeuw, D.M.d.: Low-cost all-polymer integrated circuits. *Appl. Phys. Lett.* 73, 108–110 (1998)
5. Skotheim, T.A., Reynolds, J.R.: *Conjugated Polymers: Processing and Applications*. CRC Press Taylor & Francis Group, USA (2006)
6. Brabec, C.J., Dyakonov, V., Scherf, U.: *Organic photovoltaics. Materials, device physics and manufacturing technologies*. Wiley-VCH, Weinheim (2008)
7. Schlüter, A.D.: The tenth anniversary of Suzuki polycondensation (SPC). *J. Polym. Sci. A: Polym. Chem.* 39, 1533–1556 (2001)
8. Calabrese, A., Pellegrino, A., Perin, N., Spera, S., Tacca, A., Po, R.: Optical and Electrochemical Properties of Fluorene/Thiophene/Benzothiadiazole Random Copolymers for Photovoltaic Applications (2010) (under submission)
9. Gritzner, G., Kuta, J.: Recommendations on reporting electrode potentials in nonaqueous solvents: IUPC commission on electrochemistry. *Electrochimica Acta* 29, 869–873 (1984)
10. Hehre, W.J., Radom, L., Schleyer, P.v.R., Pople, J.A.: *Ab Initio Molecular Orbital Theory*. John Wiley & Sons, Inc., USA (1986)
11. Becke, A.D.: Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A* 38, 3098–3100 (1988)
12. Hwang, S.-W., Chen, Y.: Synthesis and Electrochemical and Optical Properties of Novel Poly(aryl ether)s with Isolated Carbazole and p-Quaterphenyl Chromophores. *Macromolecules* 34, 2981–2986 (2001)
13. Johansson, T., Mammo, W., Svensson, M., Andersson, M.R., Inganäs, O.: Electrochemical bandgaps of substituted polythiophenes. *J. Mater. Chem.* 13, 1316–1323 (2003)
14. Brabec, C.J., Hummelen, J.C., Sariciftci, N.S.: Plastic Solar Cells. *Adv. Funct. Mater.* 11, 15–26 (2001)
15. Brabec, C.J., Cravino, A., Meissner, D., Sariciftci, N.S., Fromherz, T., Rispens, M.T., Sanchez, L., Hummelen, J.C.: Origin of the Open Circuit Voltage of Plastic Solar Cells. *Adv. Funct. Mater.* 11, 374–380 (2001)

Optical Transducers Based on Amorphous Si/SiC Photodiodes

Manuela Vieira^{1,2,3}, Paula Louro^{1,2}, Miguel Fernandes^{1,2}, Manuel A. Vieira^{1,2},
and João Costa^{1,2}

¹ Electronics Telecommunications and Computer Dept., ISEL, Lisbon, Portugal

² CTS-UNINOVA, Quinta da Torre, 2829-516, Caparica, Portugal

³ DEE-FCT-Universidade Nova de Lisboa, Quinta da Torre, 2829-516, Caparica, Portugal

Abstract. Amorphous Si/SiC photodiodes working as photo-sensing or wavelength sensitive devices have been widely studied. In this paper single and stacked a-SiC:H p-i-n devices, in different geometries and configurations, are reviewed. Several readout techniques, depending on the desired applications (image sensor, color sensor, wavelength division multiplexer/demultiplexer device) are proposed. Physical models are presented and supported by electrical and numerical simulations of the output characteristics of the sensors.

Keywords: Amorphous Si/SiC photodiodes, photonic, optoelectronic, image sensors, demultiplexer devices, optical amplifiers.

1 Introduction

The Tunable optical filters are useful in situations requiring spectral analysis of an optical signal. A tunable optical device is a device for wavelength selection such as an add/drop multiplexer (ADM) which enables data to enter and leave an optical network bit stream without having to demultiplex the stream. They are often used in wavelength division multiplexing (WDM) systems [1]. WDM systems have to accomplish the transient color recognition of two or more input channels in addition to their capacity of combining them onto one output signal without losing any specificity (wavelength and bit rate). Only the visible spectrum can be applied when using polymer optical fiber (POF) for communication [2]. So, the demand of new optical processing devices is a request.

2 Contribution to Sustainability

These sensors are different from the other electrically scanned image sensors as they are based on only one sensing element with an opto-mechanical readout system. No pixel architecture is needed. The advantages of this approach are quite obvious like the feasibility of large area deposition and on different substrate materials (e.g., glass, polymer foil, etc.), the simplicity of the device and associated electronics, high resolution, uniformity of measurement along the sensor and the cost/simplicity of the detector. The design allows a continuous sensor without the need for pixel-level

patterning, and so can take advantage of the amorphous silicon technology. It can also be integrated vertically, i. e. on top of a read-out electronic, which facilitates low cost large area detection systems where the signal processing can be performed by an ASIC chip underneath.

In this paper we present results on the optimization of different multilayered a-SiC:H heterostructures for spectral analysis in the visible spectrum. A theoretical analysis and an electrical simulation are performed to support the wavelength selective behavior.

3 Device Configuration and Spectral Analysis

The semiconductor sensor element is based on single or stacked a-SiC:H p-i-n structures using different architectures, as depicted in Fig. 1. All devices were produced by PE-CVD on a glass substrate. The simplest configuration is a p-i-n photodiode (NC11) where the active intrinsic layer is a double layered a-SiC:H/a-Si:H thin film. In the other the active device consists of a p-i'(a-SiC:H)-n / p-i(a-Si:H)-n heterostructures. To decrease the lateral currents, the doped layers (20 nm thick) of NC12 have low conductivities and are based on a-SiC:H [3].

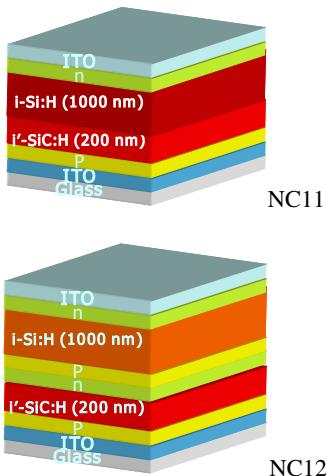


Fig. 1. Device configuration

Deposition conditions are described elsewhere [4, 5]. Full wavelength detection is achieved based on spatially separated absorption of different wavelengths. The blue sensitivity and the red transmittance were optimized, respectively, using a thin a-SiC:H front absorber (200 nm) with an optical gap of 2.1 eV and a thick a-SiH back absorber (1000 nm) with an optical gap around 1.8 eV. The thicknesses of both absorbers result from a trade-off between the full absorption of the blue light in the front diode and the green light across both. As a result, both front and back diodes act as optical filters confining, respectively, the blue and the red optical carriers, while the green ones are absorbed across both [6]. The devices were characterized through spectral response at 1 kHz and photocurrent-voltage measurements. To test the sensitivity of the device under different electrical and optical bias three modulated monochromatic

lights channels: red (R: 626 nm; $51\mu\text{W}/\text{cm}^2$), green (G: 524 nm; $73\mu\text{W}/\text{cm}^2$) and blue (B: 470nm; $115\mu\text{W}/\text{cm}^2$) and their polychromatic combinations (multiplexed signal) illuminated separately the device. The generated photocurrent was measured under positive and negative voltages ($+1\text{V} < V < -10\text{V}$), with steady state red ($\lambda_R=626 \text{ nm}; \Phi_R=102 \mu\text{W}/\text{cm}^2$), green ($\lambda_G=524 \text{ nm}; \Phi_G=71 \mu\text{W}/\text{cm}^2$) and blue ($\lambda_B=470 \text{ nm}; \Phi_B=293 \mu\text{W}/\text{cm}^2$) optical bias and without it ($\Phi_{R,G,B}=0$). The semiconductor sensor element is based on single or stacked a-SiC:H p-i-n structures using different architectures, as depicted in Fig. 1.

In Fig. 2, for NC11 and NC12 it is displayed the spectral photocurrent under different applied voltages.

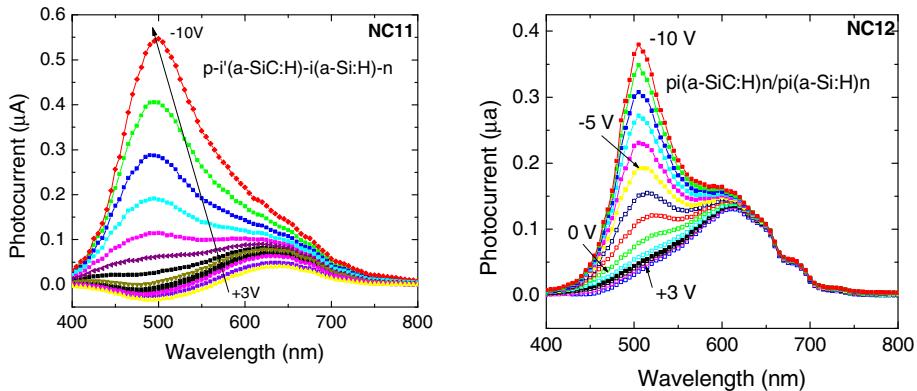


Fig. 2. Spectral response under different applied bias for three different architectures

Data show different behaviors when single and double structures are compared. In the single configuration the electrical field is asymmetrically distributed across the graded i'i-layer leading to a collection efficiency that is only dependent on the light depth penetration across it. In the double, the contribution of both front and back diodes are evident and the effect of the internal p-n junction crucial on the device functioning and future applications.

4 Wavelength (De)Multiplexer Device

In optical communications different wavelengths (color channels) which are jointly transmitted must be separated (demultiplexed) to decode the multiplexed information.

Fig. 3 displays, under positive and negative bias, the multiplexed signals, acquired with NC12 device, due to the simultaneous transmission of three independent bit sequences, each one assigned to one of the red (R: 626 nm), green (G: 524 nm) and blue (B: 470nm) color channels. At the top of the figure, the optical signal used to transmit the information is displayed to guide the eyes on the different ON-OFF states.

To recover the transmitted information (8 bit per wavelength channel, 2000bps) the multiplexed signal, during a complete cycle ($0 < t < T$), was divided into eight time slots, each corresponding to one bit where the independent optical signals can be ON (1) or OFF (0). As, under forward bias, the device has no sensitivity to the blue channel (Fig. 2), the red and green transmitted information can be identified from the multiplexed signal at +1V. The highest level corresponds to both channels ON (R&G: R=1, G=1), and the lowest to the OFF-OFF stage (R=0; G=0). The two levels in-between are related with the presence of only one channel ON, the red (R=1, G=0) or the green (R=0, G=1) (see horizontal labels in Fig. 3). To distinguish between these two situations and to decode the blue channel, the correspondent sub-levels, under

reverse bias, have to be analyzed. From Fig. 2b, it is observed that the green channel is more sensitive to changes on the applied voltage than the red, and that the blue only appears under reverse bias. So, the highest increase at -8V corresponds to the blue channel ON ($B=1$), the lowest to the ON stage of the red channel ($R=1$) and the intermediate one to the ON stage of the green ($G=1$) (vertical labels in Fig. 3). Using this simple algorithm and MATLAB as programming environment, the independent red, green and blue bit sequences can be decoded, in real time, as: R [00011011], G [001101010] and B [00101011] in agreement with the optical signals used to transmit the information.

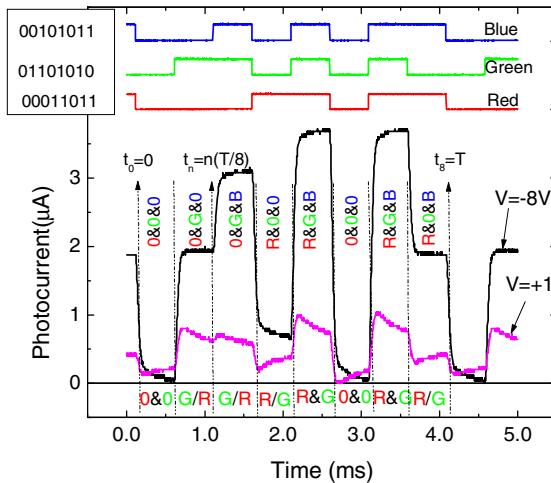


Fig. 3. Multiplexed signals under reverse and forward bias. On the top, the optical signal used to transmit the information guide the eyes on the different ON-OFF states. The recovered bit sequences are shown as an insert.

5 Optical Bias Controlled Wavelength Discrimination

In Fig. 4a the spectral photocurrent at different applied voltages is displayed without and under red, green and blue background irradiation. In Fig. 4b the ratio between the spectral photocurrents under red, green and blue steady state illumination and without it (dark) are plotted.

When an external electrical bias (forward or reverse) is applied to a double pin structure, it mainly influences the field distribution within the less photo excited sub-cell: the back under blue irradiation and the front under red steady bias [8]. Under negative bias the blue bias enhances the spectral sensitivity in the long wavelength ranges and quenches in the short wavelength range. The red bias has an opposite behavior; it reduces the collection in red/green wavelength ranges and amplifies the blue one. The green optical bias only reduces the spectral greenish photocurrent keeping the other two almost unchangeable.

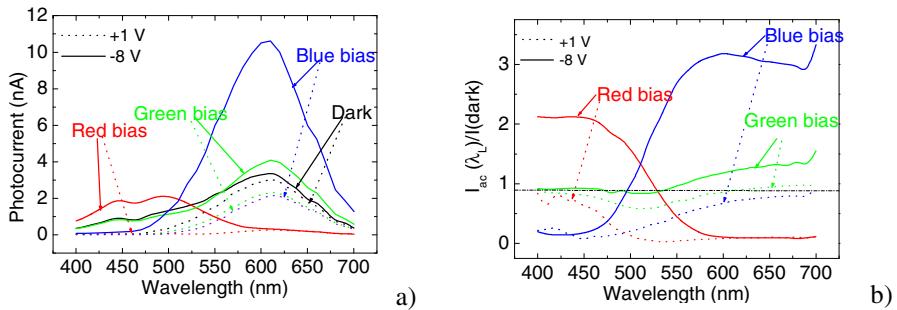


Fig. 4. a) Spectral photocurrent @ +1 V, -8 V without (dark) and under red, green and blue optical bias. b) Ratio between the photocurrents under red, green and blue steady state illumination and without it (dark).

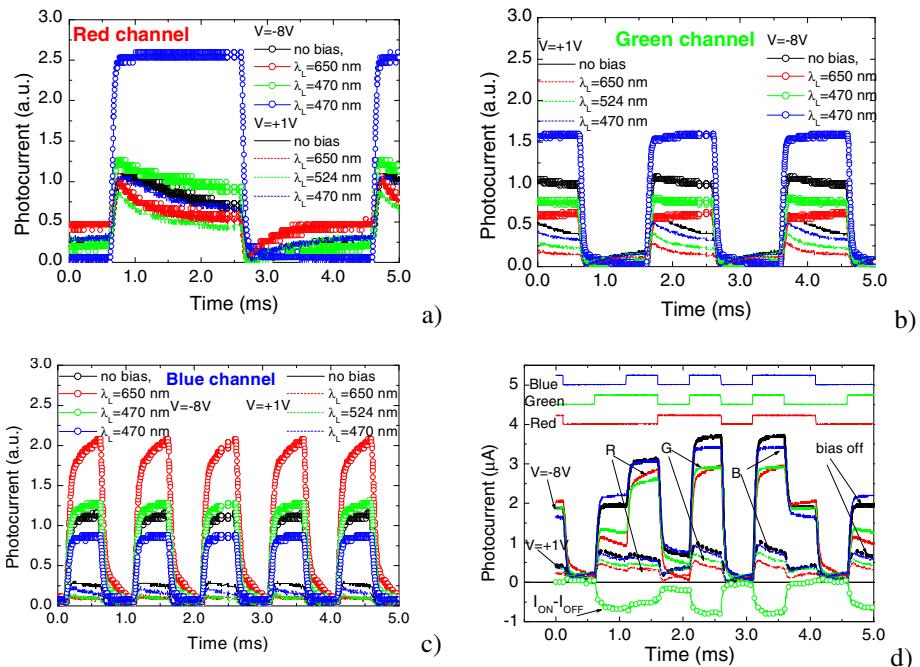


Fig. 5. Input red (a) green (b) and blue (c) signals under negative and positive bias without and with red, green and blue steady state optical bias. d) Multiplexed signals @-8V/+1V (solid /dot lines); without (bias off) and with (R, G, B) green optical bias.

To analyze the self bias amplification under transient conditions and uniform irradiation, three monochromatic pulsed lights (R, G and B input channels illuminated separately NC12 device. Steady state *red*, *green* and *blue* optical bias was superimposed separately and the photocurrent generated measured at -8V and +1 V.

In Fig. 5a, Fig. 5b and Fig. 5c the signal is displayed for each monochromatic channel separately and in Fig. 5d for the bit sequence shown on the top of figure.

Results show that when an optical bias is applied it mainly enhances the field distribution within the less photo excited sub-cell. Even at high frequencies the blue irradiation amplifies the red ($\alpha_R=2.25$) and the green ($\alpha_G=1.5$) channels and quenches the blue one ($\alpha_B=0.8$). The red bias has an opposite behavior, it reduces the red and green channels and amplifies the blue ($\alpha_R=0.9$, $\alpha_G=0.5$, $\alpha_B=2.0$). The green optical bias reduces the green channel ($\alpha_G=0.75$) keeping the other two almost unchangeable. This optical nonlinearity makes the transducer attractive for optical communications and can be used to distinguish a wavelength, to read a color image, to amplify or to suppress a color channel or to multiplex or demultiplex an optical signal. In Fig. 5c the green channel is tuned through the difference between the multiplexed signal with and without green optical bias

6 Electrical Model

The silicon-carbon pi'pnin device is a monolithic double pin photodiode structure with two red and blue optical connections for light triggering (Figure 6a). Based on the experimental results and device configuration an electrical model was developed [7]. Operation is explained in terms of the compound connected phototransistor equivalent model displayed in Figure 6b.

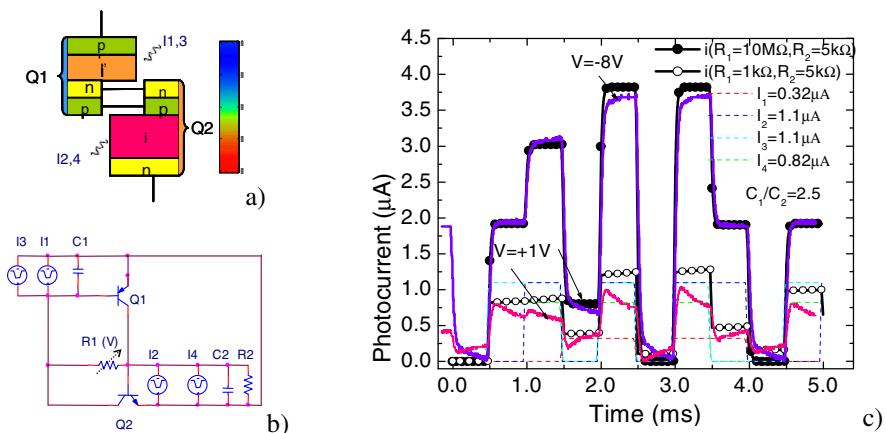


Fig. 6. a) Compound connected phototransistor equivalent model. b) ac equivalent circuit. c) Multiplexed simulated (symbols) and experimental (solid lines) results under positive and negative dc bias. The current sources used as input channels (dash lines) are displayed.

The capacitive effects due to the transient nature of the input signals are simulated through C_1 and C_2 capacitors. R_1 and R_2 model the dynamic resistance of the internal and back junctions, respectively. The photocurrent under positive (open symbols) and negative (solid symbols) dc bias is displayed in Figure 6c. We have used as input parameters the experimental values of Figure 3. The input transient current sources used to simulate the photons absorbed in the front (blue, I_1), back (red, I_2), or across

both (green, I_3 and I_4) photodiodes are also displayed (dash lines). To validate the model the experimental multiplexed signals at -8V and +1V are also shown (lines).

Good agreement between experimental and simulated data was observed. The expected levels, under reversed bias, and their reduction under forward bias are clearly seen (Figure 3). If not triggered ON by light the device is nonconducting (low levels), when turned ON it conducts through different paths depending on the applied voltage (negative or positive) and trigger connection (Q_1 , Q_2 or both).

Under negative bias (low R_1) the base emitter junction of both transistors are inversely polarized and conceived as phototransistors, thus taking advantage of the amplifier action of neighboring collector junctions, which are polarized directly. This results in a charging current gain proportional to the ratio between both collector currents (C_1/C_2). The device behaves like a transmission system able to store, amplify and transport all the minority carriers generated by the current pulses, through the capacitors C_1 and C_2 . Under positive bias (high R_1) the device remains in its non conducting state unless a light pulse (I_2 or I_2+I_4) is applied to the base of Q_2 . This pulse causes Q_2 to conduct because the reversed biased n-p internal junction behaves like a capacitor inducing a charging current across R_2 . No amplification effect was detected since Q_1 acts as a load and no charges are transferred between C_1 and C_2 .

The optical amplification under transient condition also explains the use of the same device configuration in the Laser Scanned Photodiode (LSP) image and color sensor [9]. Here, if a low power monochromatic scanner is used to readout the generated carriers the transducer recognize a color pattern projected on it acting as a color and image sensor. Scan speeds up to 10^4 lines per second are achieved without degradation in the resolution.

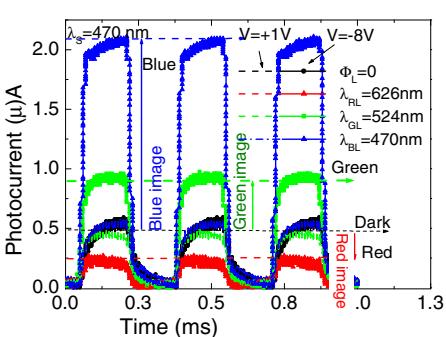


Fig. 7. LSP color sensor

For the color image sensor only the red channel is used (Fig. 6, $I_2 \neq 0$, $I_1=I_3=I_4=0$). To simulate a color image at the XY position, using the multiplexing technique, a low intensity moving red pulse scanner (Φ_S , λ_S), impinges in the device in dark or under different red, green and blue optical bias (color pattern, Φ_L , $\lambda_{RGB,L}$, $\Phi_L > \Phi_S$). Fig. 7 displays the experimental acquired electrical signals. The image signal is defined as the difference between the photocurrent with (light pattern) and without (dark) optical

bias. Without optical bias ($\Phi_L = 0$) and during the red pulse, only the minority carriers generated at the base of Q_2 by the scanner, flow across the circuit (I_2) either in reverse or forward bias. Under red irradiation (red pattern, $\Phi \neq 0$, λ_{RL}) the base-emitter junction of Q_2 is forward bias, the recombination increases reducing I_2 thus, a negative image is observed whatever the applied voltage. Under blue ($\Phi \neq 0$, λ_{BL}) or green ($\Phi \neq 0$, λ_{GL}) pattern irradiations the signal depends on the applied voltage and consequently, on R_1 . Under negative bias an optical enhancement is observed due to the amplifier action of adjacent collector junctions which are always polarized directly. Under positive bias the device remains in its non conducting state, unless the

red pulse (I_2 , dark level) is applied to the base of Q_2 . No amplification occurs and the red channel is strongly reduced when compared with its value under negative voltage. Under blue irradiation, the internal junction becomes reverse biased at +1 V (blue threshold) allowing the blue recognition. The behavior under a green pattern depends on the balance between the green absorption into the front and back diodes that determines the amount of charges stored in both capacitors. Under negative bias both the green component absorbed either in the front (blue-like) or at the back (red-like) diodes reaches the output terminal while for voltages at which the internal junction n-p becomes reversed (green threshold), the blue-like component is blocked and the red-like reduced. So, by using a thin a-SiC:H front absorber optimized for blue collection and red transmittance and a back a-Si:H absorber to spatially decouple the green/red absorption, the model explains why a moving red scanner (probe beam) can be used to readout RGB the full range of colors at each location without the use of a pixel architecture.

7 Conclusions

Different architectures based on silicon carbon pin devices were studied both theoretically and experimentally. Results show that when the doped layers has low conductivities and are based on a-SiC:H material the devices are optical and voltage controlled and can be used as optical filter, amplifier or multiplexer /demultiplexer devices in the visible range.

References

1. Bas, M.: Fiber Optics Handbook, Fiber, Devices and Systems for Optical Communication, ch. 13. McGraw-Hill, Inc., New York (2002)
2. Randel, S., Koonen, A.M.J., Lee, S.C.J., Breyer, F., Garcia Larrode, M., Yang, J., Ng’Oma, A., Rijckenberg, G.J., Boom, H.P.A.: Advanced modulation techniques for polymer optical fiber transmission. In: Proc. ECOC 2007 (Th 4.1.4), Berlin, Germany, pp. 1–4 (2007)
3. Vieira, M., Fernandes, M., Louro, P., Fantoni, A., Vygranenko, Y., Lavareda, G., Nunes de Carvalho, C.: Mat. Res. Soc. Symp. Proc., vol. 862, pp. A13.4 (2005)
4. Vieira, M., Fantoni, A., Fernandes, M., Louro, P., Lavareda, G., Carvalho, C.N.: Thin Solid Films 515(19), 7566–7570 (2007)
5. Vygranenko, Y., Wang, K., Vieira, M., Sazonov, A., Nathan, A.: Appl. Phys. Lett. 95, 263–505 (2009)
6. Louro, P., Vieira, M., Vygranenko, Y., Fantoni, A., Fernandes, M., Lavareda, G., Carvalho, N.: Mat. Res. Soc. Symp. Proc., vol. 989, p. A12.04 (2007)
7. Vieira, M.A., Vieira, M., Fernandes, M., Fantoni, A., Louro, P., Barata, M.: MRS Proceedings Amorphous and Polycrystalline Thin-Film Silicon Science and Technology 2009, vol. 1153, p. A08-0 (2009)
8. Vieira, M., Fantoni, A., Louro, P., Fernandes, M., Schwarz, R., Lavareda, G., Carvalho, C.N.: Vacuum 82(12), 1512–1516 (2008)
9. Vieira, M., Fantoni, A., Fernandes, M., Louro, P., Rodrigues, I.: Mat. Res. Soc. Symp. Proc. 762@2003 A.18.13

Author Index

- Afonso, Marcos 481
Agostinho, Carlos 45
Alcock, Jeffrey R. 3
Aliakbarpour, Hadi 189, 277
Almeida, Giovana 349
Almeida, Graça 286
Almeida, Luis 325
Álvarez, Alfredo 529, 553
Alves, Bruno 205
Andersson, Maria 277
Andrade, Aron 375, 410
Antunes, Pedro 410
Antunes, Rui 305
Aranha, André 256
Araújo, Rui Esteves 359
Ardalan, Adel 117
Arruda, Celso 410
Assadi, Amir 117
Ayanzadeh, Ramin 109

Barata, José 205
Barbosa, Paulo 237, 256
Bărbuceanu, Florin 181
Barros, João Paulo 227, 237, 256
Batista, Arnaldo 341
Bernardo, Luis 581
Bock, Eduardo 375, 410
Borza, Paul Nicolae 421, 429
Bruckner, Dietmar 197

Calabrese, Anna 596
Camaioni, Nadia 596
Camarinha-Matos, Luis M. 11, 21
Cardoso, A.J. Marques 493
Cardoso, Alberto 349, 383
Cardoso, José Roberto 375, 410
Cardoso, Tiago 21
Carp, Marius Cătălin 421, 429
Carvalho, Carlos 510
Cavalheiro, André 375, 410
Chastre, Carlos 286
Chen, Qian 591
Coito, Fernando V. 305, 393

Consoli, Alfio 596
Costa, Anikó 237, 256
Costa, Fernando 393
Costa, João 604

da Costa, João Caldas 341
Dascalu, Laura Madalina 131, 402
Dashti, Hesam 117
da Silva, Robson M. 367
Debebe, Siraye E. 596
de Castro, Ricardo 359
Dias, Jorge 165, 173, 189, 277, 325, 333
Dinis, Rui 581
Duarte-Ramos, Hermínio 305
Duguleană, Mihai 181
Durugbo, Christopher 3

Encarnação, Luis 573
Estima, Jorge O. 493

Faria, Diego R. 173
Felicetti, Alberto M. 33
Fernandes, Miguel 604
Ferreira, João 205
Figueiredo, Jorge 237, 256
Figueiredo, Michael 565
Fonseca, Jeison 375, 410
Fonseca, José 286
Fonseca, Tiago 474
Foruzantabar, Ahmad 315
Francisco, Ricardo 445

Gallardo-Lozano, Javier 502
Ganhão, Francisco 581
Gil, Paulo 383
Goes, João 565
Gomes, Henrique L. 591
Gomes, Luís 227, 237, 246, 256
Goncalves, David 45
González, Pedro 466
González Romera, Eva 457, 466
Grzech, Adam 75

- Guerrero, Miguel A. 466
 Gzara, Lilia 67
- Hachani, Safa 67
 Hadjinicolaou, M. 297
- Inácio, David 529, 545
 Ionescu, Razvan Mihai 518
- Jardim-Goncalves, Ricardo 45
 Jerbić, Bojan 147
 Josué, João Gil 437
 Junqueira, Fabrício 57
 Juszczyszyn, Krzysztof 75
- Karapidakis, E. 297
 Katsioulis, V. 297
 Khoshhal, Kamrad 189, 277
 Kiazadeh, Asal 591
- Lavareda, Guilherme 510
 Leão, Tarcísio 410
 Leme, Juliana 410
 León-Sánchez, Ana Isabel 502
 Lima, José 537
 Lobo, Jorge 173, 215
 Longo, Luca 596
 Lopes, Gabriel 101
 Louro, Paula 604
 Luís, Gonçalo F. 545
- Macedo, Mário 581
 Macedo, Patrícia 11
 Marques, José 349
 Martins, João F. 445, 474, 481, 529
 Martins, Ricardo 173
 Mekhnacha, Kamel 189, 277
 Melicio, Fernando 286
 Menezes, Paulo 325
 Miguel, Ignacio de 153
 Milanés-Montero, María Isabel 457, 502
 Mirzaee, Alireza 315
 Miyagi, Paulo Eigi 57, 367, 375
 Mogan, Gheorghe Leonte 269
 Monteiro, André 237
 Moutinho, Filipe 237, 246, 256
- Nedelcu, Adrian 181
 Negoita, Andrei 518
- Neves, Mário Ventim 437, 529, 537, 545, 553
 Nunes, Gonçalo 383
- Oliveira, Luís B. 565
 Oliveira, Rodolfo 581
 Ortigueira, Manuel Duarte 341
- Pais, Rui 227
 Palma, Luís 393
 Patanè, Salvatore 596
 Paulino, Nuno 510
 Pellegrino, Andrea 596
 Pereira, Fernando 246
 Pereira, Miguel 581
 Pereira, Paulo 581
 Pereira, Pedro 445, 481
 Perin, Nicola 596
 Pessoa, Marcosiris A.O. 57
 Peter, Ioan 518
 Pina, João Murta 437, 545, 553
 Pinto, Paulo 581
 Pinto, Sónia F. 573
 Po, Riccardo 596
 Portugal, David 139
 Postelnicu, Cristian-Cezar 157
 Prado, José Augusto 165
 Pronto, Anabela 537
 Prusiewicz, Agnieszka 83
 Pușcaș, Ana Maria 421, 429
- Quintas, João 189, 277
- Ramalho, Franklin 237, 256
 Redondo, Luis. M. 573
 Ribeiro, Luis 205
 Ribeiro, Rita A. 101, 109
 Rocha, Paulo R. 591
 Rocha, Rui 139
 Rodrigues, Amadeu Leão 529, 553
 Romero-Cadaval, Enrique 457, 466, 502
 Ros, Julien 189, 277
 Ruiz Arranz, Sergio 457
- Santin, Edinei 565
 Santos, Amâncio 383
 Santos, Luís 333
 Santos Filho, Diolino J. 57, 367, 375, 410
 Sarraipa, João 45

- Scutaru, Gheorghe 518
Šekoranja, Bojan 147
Selen, Ebru Selin 117
Setayeshi, Saeid 109
Shahamatnia, Ehsan 109
Silva, Cibele 410
Silva, José Fernando 573
Silveira, Alexandre 359
Simplício, Carlos 165
Stavar, Adrian 131, 402
Stelmach, Paweł 75
Švaco, Marko 147
Świątek, Jerzy 91
Szekely, Iuliu 421
Tacca, Alessandra 596
Talaat, Adel 117
Talaba, Doru 131, 157, 402
Teixeira, Luís 101
Tinti, Francesca 596
Tiwari, Ashutosh 3
Toma, Madalina-Ioana 157
Tomczak, Jakub M. 91
Topoleanu, Tudor-Sabin 269
Trindade, Pedro 215
Tsikalakis, A. 297
Uebelhart, Beatriz 410
Utiyama, Bruno 410
Verjus, Hervé 67
Vieira, José 349
Vieira, Manuel A. 604
Vieira, Manuela 604
Vlad, Stoianovici 181
Volpentesta, Antonio P. 33
Yin, GuoQing 197
Zięba, Maciej 83