

Lecture Notes in Networks and Systems 708

Marek Pawelczyk
Dariusz Bismor
Szymon Ogonowski
Janusz Kacprzyk *Editors*

Advanced, Contemporary Control

Proceedings of the XXI Polish Control
Conference, Gliwice, Poland, 2023.
Volume 1

 Springer

Series Editor

Janusz Kacprzyk, *Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland*

Advisory Editors

Fernando Gomide, *Department of Computer Engineering and Automation—DCA, School of Electrical and Computer Engineering—FEEC, University of Campinas—UNICAMP, São Paulo, Brazil*

Okay Kaynak, *Department of Electrical and Electronic Engineering, Bogazici University, Istanbul, Türkiye*

Derong Liu, *Department of Electrical and Computer Engineering, University of Illinois at Chicago, Chicago, USA*

Institute of Automation, Chinese Academy of Sciences, Beijing, China

Witold Pedrycz, *Department of Electrical and Computer Engineering, University of Alberta, Alberta, Canada*

Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland

Marios M. Polycarpou, *Department of Electrical and Computer Engineering, KIOS Research Center for Intelligent Systems and Networks, University of Cyprus, Nicosia, Cyprus*

Imre J. Rudas, *Óbuda University, Budapest, Hungary*

Jun Wang, *Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong*

The series “Lecture Notes in Networks and Systems” publishes the latest developments in Networks and Systems—quickly, informally and with high quality. Original research reported in proceedings and post-proceedings represents the core of LNNS.

Volumes published in LNNS embrace all aspects and subfields of, as well as new challenges in, Networks and Systems.

The series contains proceedings and edited volumes in systems and networks, spanning the areas of Cyber-Physical Systems, Autonomous Systems, Sensor Networks, Control Systems, Energy Systems, Automotive Systems, Biological Systems, Vehicular Networking and Connected Vehicles, Aerospace Systems, Automation, Manufacturing, Smart Grids, Nonlinear Systems, Power Systems, Robotics, Social Systems, Economic Systems and other. Of particular value to both the contributors and the readership are the short publication timeframe and the world-wide distribution and exposure which enable both a wide and rapid dissemination of research output.

The series covers the theory, applications, and perspectives on the state of the art and future developments relevant to systems and networks, decision making, control, complex processes and related areas, as embedded in the fields of interdisciplinary and applied sciences, engineering, computer science, physics, economics, social, and life sciences, as well as the paradigms and methodologies behind them.

Indexed by SCOPUS, INSPEC, WTI Frankfurt eG, zbMATH, SCImago.

All books published in the series are submitted for consideration in Web of Science.

For proposals from Asia please contact Aninda Bose (aninda.bose@springer.com).

Marek Pawelczyk · Dariusz Bismor ·
Szymon Ogonowski · Janusz Kacprzyk
Editors

Advanced, Contemporary Control

Proceedings of the XXI Polish Control
Conference, Gliwice, Poland, 2023. Volume 1

 Springer

 **PCC2023**
Polish Control Conference

Editors

Marek Pawelczyk
Silesian University of Technology
Gliwice, Poland

Dariusz Bismor
Silesian University of Technology
Gliwice, Poland

Szymon Ogonowski
Silesian University of Technology
Gliwice, Poland

Janusz Kacprzyk
Systems Research Institute
Polish Academy of Science
Warsaw, Poland

ISSN 2367-3370

ISSN 2367-3389 (electronic)

Lecture Notes in Networks and Systems

ISBN 978-3-031-35169-3

ISBN 978-3-031-35170-9 (eBook)

<https://doi.org/10.1007/978-3-031-35170-9>

© The Editor(s) (if applicable) and The Author(s), under exclusive license
to Springer Nature Switzerland AG 2023

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, 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.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

This book is a collection of contributed works concerned with many aspects of modern control sciences and robotics. The works were submitted to the XXI Polish Control Conference, which is a triennial meeting organized under the substantive supervision of the Automation and Robotics Committee of the Polish Academy of Sciences since 1958. The program committee of the conference is made up of members of the Automatics and Robotics Committee of the Polish Academy of Sciences, who, by discussion and voting, entrust the organization of the conference to selected renowned academic centers in the country. In 2023, the honor of organizing and hosting the conference was given to the Silesian University of Technology, which is a winner in the Initiative of Excellence—Research University competition. The university, among its six priority research areas declared in the development strategy, focuses two of them on issues directly related to the subject of the conference.

The main purpose of the conference, in addition to integrating a wide environment of control sciences and robotics, people related to data processing and analysis, mathematical modeling, artificial intelligence methods and many other related areas, is to present the latest achievements and exchange experiences in the field of research and application work. It is the most important scientific forum in this field organized in Poland, hosting also a selection of excellent guests from abroad. The organizers have also extended their hospitality to guests from Ukraine, for which a special support was provided, with the help of the Polish Ministry of Education and Science.

The organizing committee received 88 submissions. A careful peer-reviewed process, during which at least two independent reviewers were involved for each paper, allowed to select 81 papers with confirmed quality, suitable for this publication. Among this number, there are nine papers presenting the results of Project-Based Learning, in preparation of which participated students of the Silesian University of Technology.

Judging by the submissions, the most common topic in current control sciences research is modeling, identification and analysis of automation systems. However, design of control systems and fault diagnosis and fault-tolerant control are also very hot topics. A number of authors were also inspired by the statistical and stochastic methods in control engineering applications. Quite a few participants were interested in new applications of artificial intelligence and machine learning in automated and connected vehicles, in biological, medical and ecological systems (from the control sciences point of view) and in optimization and quantum computing. Several authors selected mechatronics and robotics as the topic of their submissions. Design and control of autonomous marine, robotics and vehicles systems also attracted attention of a number of researchers. Finally, several submissions present usage of control principles in other fields of research and development, including teaching and social sciences, and the organizing committee do appreciate those interdisciplinary approaches.

The editors of this volume would like to thank all the authors of the chapters included in this volume. High quality of their work greatly contributed to originality and excellence of this book. The editors also thank for the joint effort of the reviewers, who accepted the invitation to prepare the review in relatively short time, despite their other responsibilities. We particularly appreciate the reviewers who prepared three or more reviews, (in the alphabetic order): Professors Artur Babiarez, Sebastian Budzan, Witold Byrski, Jacek Czeczot, Jerzy Kasprzyk, Józef Korbicz, Zdzisław Kowalczyk, dr Zbigniew Ogonowski and professors Krzysztof Stebel and Józef Wiora. Thanks to them, the conference repeated the success of its previous editions and delivered a large amount of new and novel knowledge to its participants.

April 2023

Dariusz Bismor

Contents

Modeling, Identification, and Analysis of Automation Systems

Development of Information Technologies for the Research of Technical Systems	3
<i>S. Y. Liaskovska and Y. V. Martyn</i>	
On Mikhailov Stability Conditions for a Class of Integer- and Commensurate Fractional-Order Discrete-Time Systems	16
<i>Rafał Stanisławski and Marek Rydel</i>	
Verification of a Building Simulator in Real Experiments	27
<i>Karol Jabłoński and Dariusz Bismor</i>	
Aspects of Measurement Data Acquisition and Optimisation in the Energy Transformation of Industrial Facilities	39
<i>Lukasz Korus and Andrzej Jabłoński</i>	
Continuous-Time Dynamic Model Identification Using Binary-Valued Observations of Input and Output Signals	49
<i>Jarosław Figwer</i>	
Time Series Identification Using Monte Carlo Method	59
<i>Teresa Główka</i>	
Calculation Method of the Centrifugal Pump Flow Rate Based on Its Nominal Data and Pump Head Increase	69
<i>Uliana Nykolyn and Petro Nykolyn</i>	
A Majorization-Minimization Algorithm for Optimal Sensor Location in Distributed Parameter Systems	76
<i>Dariusz Uciński</i>	
Earned Value Method in Public Project Monitoring	86
<i>V. M. Molokanova</i>	
A Tube-Based MPC Structure for Fractional-Order Systems	103
<i>Stefan Domek</i>	
Creep Testing Machine Identification for Power System Load Optimization	113
<i>Michał Szulc, Jerzy Kasprzyk, and Jacek Loska</i>	

Development and Testing of the RFID Gripper Prototype for the Astorino Didactic Robot	123
<i>Adrian Kampa, Krzysztof Foit, Agnieszka Sekala, Jakub Kulik, Krzysztof Łukowicz, Miłosz Mróz, Julia Nowak, Marek Witański, Patryk Żebrowski, Tomasz Błaszczyk, and Dariusz Rodzik</i>	
Development of an “Artificial Lung” System for Use in Indoor Air Quality Testing	135
<i>Andrzej Kozyra, Aleksandra Lipczyńska, Piotr Koper, Radosław Babisz, Damian Madej, Konrad Nowakowski, Jakub Karwatka, and Dominik Tomczok</i>	
Recent Advances in Artificial Autonomous Decision Systems and Their Applications	145
<i>Andrzej M. J. Skulimowski, Inez Badecka, Masoud Karimi, Paweł Łydek, and Przemysław Pukocz</i>	
Modeling of Thermal Processes in a Microcontroller System with the Use of Hybrid, Fractional Order Transfer Functions	158
<i>Krzysztof Oprędkiewicz, Maciej Rosół, and Wojciech Mitkowski</i>	
Identification of Magnetorheological Damper Model for Off-Road Vehicle Suspension	173
<i>Piotr Krauze, Marek Płaczek, Zbigniew Żmudka, Dawid Bauke, Przemysław Olszówka, Jakub Turek, Artur Wyciśłok, Maciej Ziaja, Szymon Zosgórnik, Wojciech Janusz, Grzegorz Przybyła, and Michał Wychowański</i>	
Physics-Informed Hybrid Neural Network Model for MPC: A Fuzzy Approach	183
<i>Krzysztof Zarzycki and Maciej Ławryńczuk</i>	
Evaluating Hydrodynamic Indices of the Underground Gas Storage Operation Based upon a Two-Phase Filtration Model	193
<i>Ivan Sadovenko, Olexander Inkin, and Nataliia Dereviahina</i>	
New Metrics of Fault Distinguishability	205
<i>Jan Maciej Kościelny and Michał Bartyś</i>	
Active Noise Control with Passive Error Signal Shaping - A Critical Case Study	216
<i>Małgorzata I. Michalczyk</i>	

Fault Diagnosis and Fault-Tolerant Control

Neural Network Based Active Fault Diagnosis with a Statistical Test	227
<i>Ivo Punčochář and Ladislav Král</i>	
Fault-Tolerant Fast Power Tracking Control for Wind Turbines Above Rated Wind Speed	237
<i>Horst Schulte and Nico Goldschmidt</i>	
A Multiple Actuator and Sensor Fault Estimation for Dynamic Systems	250
<i>Marcin Pazera, Marcin Witczak, and Józef Korbicz</i>	
Enhancing Power Generation Efficiency of Piezoelectric Energy Harvesting Systems: A Performance Analysis	261
<i>Bartłomiej Ambrożkiewicz, Zbigniew Czyż, Paweł Stączek, Jakub Anczarski, and Mikołaj Jachowicz</i>	
Power Quality Issues of Photovoltaic Stations in Electric Grids and Control of Main Parameters Electromagnetic Compatibility	269
<i>Yaroslav Batsala and Ivan Hlad</i>	
Integration of Fault-Tolerant Design and Fault-Tolerant Control of Automated Guided Vehicles	277
<i>Ralf Stetter and Marcin Witczak</i>	
Problems of Using Eddy Current Arrays NDT	287
<i>Iuliia Lysenko, Yuriy Kuts, Valentyn Uchanin, Yordan Mirchev, and Alexander Alexiev</i>	
Calibration of a High Sampling Frequency MEMS-Based Vibration Measurement System	294
<i>Muhammad Ahsan and Dariusz Bismor</i>	

Design of Control Systems

Configurable Dynamics of Electromagnetic Suspension by Fuzzy Takagi-Sugeno Controller	305
<i>Adam Krzysztof Pilat, Hubert Milanowski, Rafal Bieszczad, and Bartłomiej Sikora</i>	
Consistent Design of PID Controllers for Time-Delay Plants	320
<i>Andrzej Bożek, Zbigniew Świder, and Leszek Trybus</i>	

Synchronization of Four Different Chaotic Communication Systems with the Aim of Secure Communication	329
<i>Ali Soltani Sharif Abadi, Pooyan Alinaghi Hosseinabadi, and Andrew Ordys</i>	
Design of Robust H_∞ Control of an Active Inerter-Based Vehicle Suspension System	337
<i>Keyvan Karim Afshar, Roman Korzeniowski, and Jarosław Konieczny</i>	
Nonlinear Adaptive Control with Invertible Fuzzy Model	349
<i>Marcin Jastrzębski, Jacek Kabziński, and Rafał Zawisław</i>	
On the Choice of the Cost Function for Nonlinear Model Predictive Control: A Multi-criteria Evaluation	361
<i>Robert Nebeluk and Maciej Ławryńczuk</i>	
Adaptive Sliding Mode control of Traffic Flow in Uncertain Urban Networks	372
<i>Ali Soltani Sharif Abadi, Pooyan Alinaghi Hosseinabadi, and Andrew Ordys</i>	
An Application of the Dynamic Decoupling Techniques for a Nonlinear TITO Plant	381
<i>Szymon Król and Paweł Dworak</i>	
Attitude Control of an Earth Observation Satellite with a Solar Panel	393
<i>Zbigniew Emirsajłow, Tomasz Barciński, and Nikola Bukowiecka</i>	
Author Index	403

Modeling, Identification, and Analysis of Automation Systems



Development of Information Technologies for the Research of Technical Systems

S. Y. Liaskovska¹(✉) and Y. V. Martyn²

¹ Department of Artificial Intelligence, Lviv Polytechnic National University,
Kniazia Romana Str., 5, Lviv 79905, Ukraine
solomiam@gmail.com

² Department of Project Management, Information Technologies and Telecommunications,
Lviv State University of Life Safety, Kleparivska Str., 35, Lviv 79007, Ukraine

Abstract. Information technology is rapidly developing and finding practical applications in various fields of human activity. Relevant is the study of technical systems. The final functional action of most of them is the delivery of a formed signal and, accordingly, a mechanical action on the object. The study consists of two parts: information technology and an executive mechanism. These components combine two main conditions: stability and functionality in operation. A rational choice of the parameters of the components is a common requirement for ensuring their stability and functionality. Modeling is used to fulfill this requirement [1–3]. In the process of studying technical systems using the created model, the interaction of various parameters is taken into account, which determines the overall stability of the operation and the efficiency of the given mode of operation of the system. It should be noted that solutions to most problems are visualized by graphical dependencies. Visualization and study of complex and simple functional dependencies in multidimensional spaces of models of multiparametric systems, which include technical systems, are reduced to formalized geometric operations on a variety of surfaces and hypersurfaces that represent geometric models of these dependencies [4–6]. The nonlinearity of the parameters that are part of the coefficients of various equations requires the use of tools that are effective for studying a particular class of systems.

Keywords: information technology · phase · multiparameter dependencies · calculations

1 Introduction

When investigating objects, systems, and processes with many variables represented by numbers of different dimensions and parameters, tools based on geometric interpretation of functional relationships between these variables are used. These tools differ in their geometric visualization and imagery of the model, and therefore in understanding the nature of the interaction of its individual elements. Geometric models are implemented as certain geometric figures in an encompassing multidimensional space. The dimension of such a space is determined by the number of variable parameters and the dimensionality

of the numbers that represent its measures. To solve a number of practical problems related to the study and design of multiparameter technical systems, coordinate systems of spaces forming hyper-surfaces and polytopes are used as geometric models of the investigated multiparameter dependencies [6, 10–13].

2 Geometric Foundations of Model Development

Let's consider the features of forming multi-type Euclidean n -dimensional spaces as the number of independent variable parameters grows. We will fix the value of the variable parameter x , for example, $x = A$ (Fig. 1).

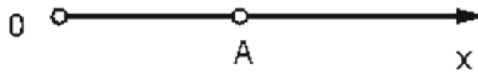


Fig. 1. The geometric object of the one-dimensional space Ox .

The manifold that reflects such a dependence belongs to the number axis Ox at point A . Since its position is determined by the coordinate $x = A$, the dimensionality of such a manifold is equal to zero. Therefore, we have a point, a zero-dimensional object of a one-dimensional space - the number axis Ox .

Let's write the mentioned equation as follows:

$$x - A = 0 \quad (1)$$

In the right-hand side, we have a fixed number, in particular, zero. Taking it as a partial value of the function, for example, y , we rewrite this equation as follows:

$$x - A = y \quad (2)$$

In the orthogonal coordinate system Oxy , such a functional dependence is represented by a line segment (Fig. 2). The dimensionality of the obtained manifold - a straight line - is equal to one, since to determine the position of any of its points, for example, point B , it is necessary to fix either $x = x_1$ or $y = y_1$. By fixing $x = A$, we obtain a zero-dimensional manifold, that is, a point of the number axis Ox as a one-dimensional subspace of the two-dimensional space (two-dimensional plane) Oxy .

Let's represent (2) as follows:

$$x - A - y = 0 \quad (3)$$

and let's generalize it for an arbitrary variable parameter z :

$$x - A - y = z \quad (4)$$

In the orthogonal coordinate system $Oxyz$, the geometric model of Eq. (4) is a plane that intersects the coordinate planes Oxy , Oxz , Oyz , forming a triangle of traces (see Fig. 3). The dimensionality of the obtained manifold - a plane - is two: the position of

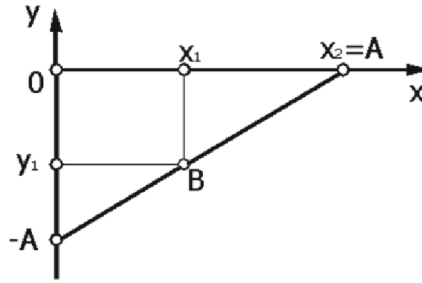


Fig. 2. Geometric model of linear dependence.

any of its points, for example, point B , is determined by two parameters at once, such as $x = x_1$, $y = y_1$ which determine the third parameter z_1 , it depends on the two listed parameters. Such a manifold - a plane - is also called a two-dimensional surface, 2 - surface or hyperplane of the three-dimensional Euclidean space E^3 .

From Fig. 3, we can see that the plane is defined by three lines that lie in the coordinate plane. Each of these lines is a one-dimensional object in the two-dimensional space. By fixing the value of one of the coordinates, for example, $x_2 = A$, we obtain a zero-dimensional object - a point - of a one-dimensional subspace, namely, the number axis Ox , Oy , or Oz (see Fig. 2).

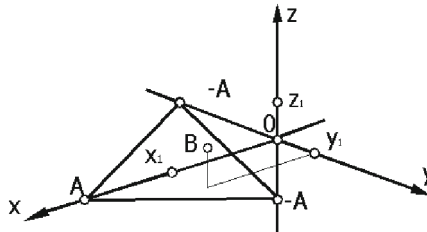


Fig. 3. Geometric model of a plane in three-dimensional space $Oxyz$.

We obtain a line as the intersection of the plane with a coordinate plane, for example, with the Oxy plane. In turn, each of the Ox , Oy , or Oz axes has zero-dimensional objects - points - of one-dimensional space as components of one-dimensional and two-dimensional objects.

We can express Eq. (4) as follows:

$$x - y - z - A = 0 \quad (5)$$

and generalize it for an arbitrary parameter t :

$$x - y - z - A = t \quad (6)$$

In the orthogonal four-dimensional coordinate system $Oxyzt$ (see Fig. 4), Eq. (6) is represented by a three-dimensional plane - a hyperplane, since the position of any point on it, for example, point B , is determined by three independent parameters, for instance, $x = x_1$, $y = y_1$, $z = z_1$, and t , which determine the fourth dependent parameter.

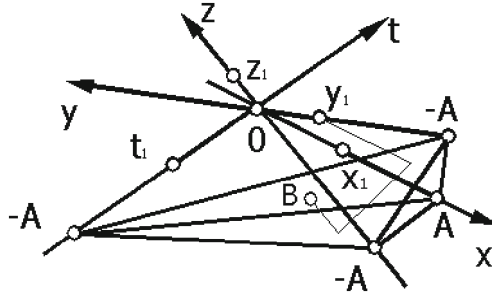


Fig. 4. Geometric model of a hyperplane in four-dimensional space $Oxyzt$.

3 Applications of Coordinate Spaces a Three-Dimensional E^3 and K^3 Space Can Serve as “Parts” of a Model

The diversity of practical problems in the analysis of multiparameter technical systems is based on the use of elementary building blocks and geometric shapes. These building blocks include points and lines. Sets of points that form lines are represented by integral curves and phase trajectories in multi-dimensional spaces. For example, they can be related by analytic dependence:

$$\begin{aligned}
 x_1 &= x_1(z); \\
 x_2 &= x_2(z); \\
 &\dots \\
 x_n &= x_n(z)
 \end{aligned}
 \tag{7}$$

n independent variables x_1, x_2, \dots, x_n that simultaneously depend on one variable z are graphically represented by curves on two-dimensional planes $x_i z$. In a multidimensional space $Oz x_1, x_2, \dots, x_n$, the geometric model (7) is a multidimensional curve line: its projection in $(n-1)$ -dimensional subspace Ox_1, x_2, \dots, x_n is also a curve line, which represents an integral curve of the process. By eliminating the argument in (7), we obtain a phase trajectory. As we can see, this approach allows reducing the number of curve lines and thus the complexity of space. Such a space is called an n -dimensional subspace or a phase space of the system [12, 14, 15].

One-dimensional n -spatial lines, l -manifolds, are the basic geometric primitive in constructing manifolds and higher-dimensional hypersurfaces, as well as boundaries of parameter spaces of multiparameter technical systems that determine their similar properties. A three-dimensional curve line in Euclidean space is defined by the equations of two surfaces that intersect:

$$\begin{aligned}
 z_1 &= z_1(x, y); \\
 z_2 &= z_2(x, y).
 \end{aligned}
 \tag{8}$$

The projections of a curve in two-dimensional planes are lines obtained by using cylinders that project these projections with respect to these planes; the curve line simultaneously belongs to the generators of each of the three cylinders. Therefore, the surfaces

(8) are used only for shaping curve lines. In a particular case, each of the surfaces can be a projection with respect to the projection planes cylinder. Such a cylinder is defined by the equation of the generator in the corresponding projection plane, for example:

$$\begin{aligned} z &= z(x); \\ y &= y(x). \end{aligned} \quad (9)$$

Then the curve is defined by the intersection of two cylinders. Note that the generators of each of the cylinders are straight lines, one-dimensional manifolds, parallel to the axis absent in the mentioned projection plane. A particular case of surfaces (8) and directionals (9) are planes and line segments that define a spatial straight line. It should be noted that the given Eqs. (8) and (9) determine the unique ways of representing curve lines in three-dimensional space.

In the four-dimensional Euclidean space $Oxyzt$, a four-dimensional line, a l -manifold in 4D, is defined as the intersection of the geometric objects in the form of a hyper-surface of this space and a two-dimensional surface, for example:

$$\begin{aligned} z_1 &= z_1(x, y, t); \\ z_2 &= z_2(x, y). \end{aligned} \quad (10)$$

The hyper-surface of the four-dimensional space has a dimensionality of $l = 3$ and its intersection with a two-dimensional plane ($m = 2$) in the four-dimensional space ($n = 4$) defines a one-dimensional ($r = 1$) curve.

$$r = l + m - n = 1. \quad (11)$$

when intersecting hyper-surfaces of a three-dimensional space with two-dimensional surfaces, we obtain a three-dimensional curve line according to (8). For a four-dimensional space, the dimensionality of a hyper-surface is one less than that of the space itself. Therefore, for two hyper-surfaces, we obtain a two-dimensional surface curve.

$$\begin{aligned} z_1 &= z_1(x, y, t); \\ z_2 &= z_2(x, y, t), \end{aligned} \quad (12)$$

result of the intersection is a two-dimensional surface:

$$r = m_1 + m_2 - n = 2. \quad (13)$$

From the analysis of (13), we have an increase in the dimensionality of the manifold intersection of two hyper-surfaces in n -space with a corresponding increase in its dimensionality.

4 The Formation of the Curve in the Four-Dimensional Space $Oxyzt$

Graphically, a line is defined as the geometric locus of points that satisfy condition (2). The projections of a multi-dimensional line are one-dimensional two-dimensional curves in the coordinate two-dimensional projection planes. Such projections are considered as the generatrices of multi-dimensional cylinders, the dimensionality of which is determined by the dimensionality of the space.

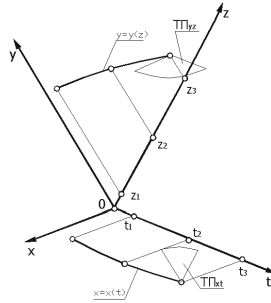


Fig. 6. Presentation 1 - of a multi-faceted four-dimensional space Oxyzt with two projections.

The components (15) define the positions of the directions $x = x(t)$ and $y = y(z)$ in the projection planes Oxt and Oyz , respectively (Fig. 6).

The formation of projection images of n -dimensional spaces involves the use of linear subspaces of dimension $n-1$. For example, projections of geometric objects in three-dimensional Euclidean space are represented on two two-dimensional planes $n-1 = 3-1 = 2$. This sufficient number of planes includes all three dimensions of the space for the three coordinate two-dimensional planes. For four-dimensional space, the minimum number of two-dimensional projection planes is two, with four dimensions of this space included. The minimum number of three-dimensional planes is also two, with two repeated dimensions. Figure 7 shows a Schoute diagram for l -forms of Euclidean (Fig. 7a) and complex (Fig. 7b) spaces.

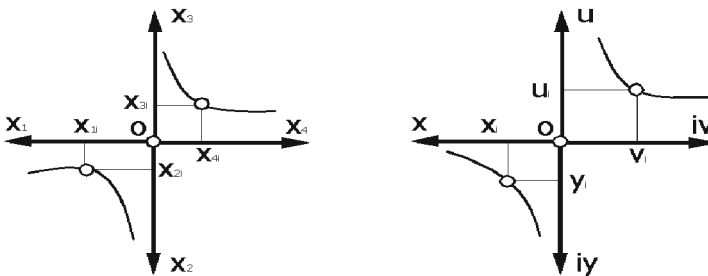


Fig. 7. The figures (a) and (b) show the Schoute diagrams for 1 - multiforms of the Euclidean 4-space and complex space, respectively.

For the Euclidean 4-space, the 1-form can be represented by a system of two equations (Fig. 7, a).

$$\begin{aligned} x_2 &= x_2(x_1); \\ x_3 &= x_3(x_4). \end{aligned} \tag{16}$$

A 1 - multiform as a graphical representation of a complex function of a real variable can be obtained for a known dependence between the components of a complex

argument, for example $y = y(x)$:

$$\omega = \omega(z) = z(x + iy(x)) = u(x, y(x)) + iv(x, y(x)) = u(x) + iv(x). \quad (17)$$

By excluding the real component of the complex argument in the right-hand side of (17), we obtain a relationship between the real and imaginary components of the complex real variable.

$$u = u(v), \quad (18)$$

This is realized by a graphical dependence in the plane of images of the function. From Fig. 7b, we have that for any $x = x_i$, we obtain, according to (17) and (18), the points whose values of y_i, u_i, v_i and determine the projections of the complex function of a real variable. From Fig. 8a, we have that both dependencies (16) uniquely determine the position of the multi-faceted object in the four-dimensional Euclidean space. However, analyzing it, we notice the absence of a connection between the two projections: given the known value of the variable x_i , it is impossible to determine the points of the second projection of the multi-faceted object. Additional conditions are required to determine the position of the multi-faceted object. Regarding Fig. 6a, one of the dependencies should be additionally presented.

$$x_3 = x_3(x_1), \quad (19)$$

or

$$x_2 = x_2(x_4). \quad (20)$$

In this case, the definition of the multi-faceted object does not exceed the bounds of the Euclidean 4-space. The multi-faceted object (16) can also be expressed in a parametric form.

$$\begin{aligned} x_1 &= x_1(t); \\ x_2 &= x_2(t); \\ x_3 &= x_3(t); \\ x_4 &= x_4(t), \end{aligned} \quad (21)$$

Transient processes when the parameters of systems, processes and objects change are represented by differential equations. A first-order differential equation describes a change in one parameter, for example, x

$$\frac{dx}{dt} = f(t, x). \quad (22)$$

The solution (22) is given by the integral curve of the two-dimensional plane Oxt . Such a curve can be constructed by numerical methods according to (11) or represented by an equation during its analytical solution. Note that the plane of integral curves often occupies the left or right part of the Oxt plane in the case of technical systems (Fig. 8 a).

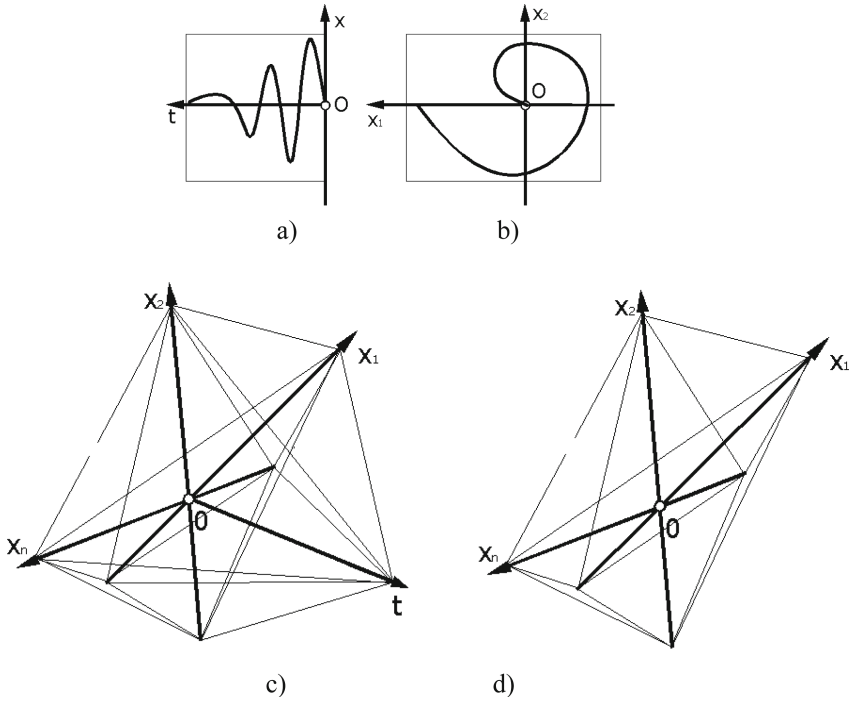


Fig. 8. Areas of integral curves and phase trajectories of multiparameter technical systems

Changing two parameters, for example, x_1 and x_2 determines the differential equation of this system of the second order, which can be given with respect to one variable, for example, x_1

$$\frac{d^2x_1}{dt^2} = f(t, x_1, \frac{dx_1}{dt}) \quad (23)$$

or systems of two differential equations of the first order

$$\begin{aligned} \frac{dx_1}{dt} &= f(t, x_1, x_2); \\ \frac{dx_2}{dt} &= \phi(t, x_1, x_2). \end{aligned} \quad (24)$$

We present the transient process (23) in the form:

$$\frac{d^3y}{dt^3} + \frac{d^2y}{dt^2} + \frac{dy}{dt} = 0, \quad (25)$$

by decreasing the order of which we form a system of three equations

$$\begin{aligned}
 y_1 &= y, \quad y_2 = \frac{dy}{dt}, \quad y_3 = \frac{d^2y}{dt^2}; \\
 \frac{dy_1}{dt} &= y_2; \quad \frac{dy_2}{dt} = y_3; \\
 \frac{dy_3}{dt} &= \frac{d^3y}{dt^3} = -y_3 - y_2.
 \end{aligned} \tag{26}$$

Solution program (26) for three-dimensional space.

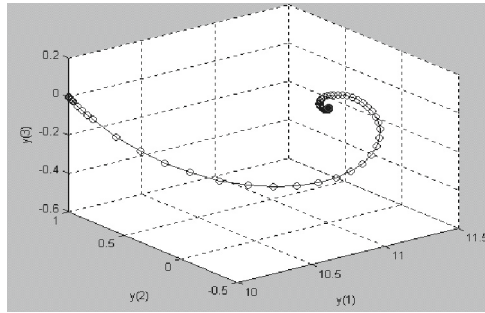


Fig. 9. The three-dimensional phase space of the system

The developed tools allow also to obtain the solution (26) as projections in two-dimensional phase planes (Fig. 10).

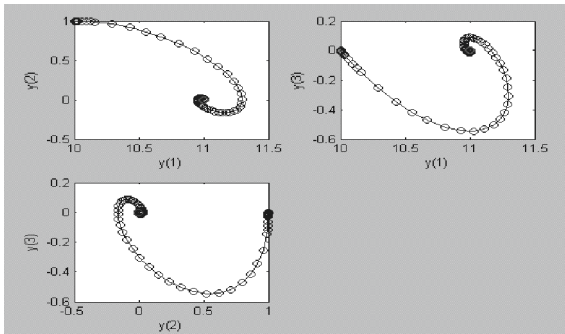


Fig. 10. Projections of trajectories in all two-dimensional phase planes

5 The Results

In the technical system, there is an interrelationship of the main elements: the object of regulation is affected simultaneously by disturbing and regulating actions, the control system allows maintaining the parameters in accordance with the technological process (Fig. 8).

The phase trajectory can be presented with the involvement of developed epurs of multidimensional spaces both on all three-dimensional and two-dimensional projection planes. In control systems of multiparameter technical objects, however, there is often a need to analyze processes involving interrelated not only two or three variable parameters (often more), but also their combinations. Matlab tools make it possible to obtain combinations of phase trajectory projections in three-dimensional and two-dimensional (or several two-dimensional) projection planes, as well as in two-dimensional and three-dimensional (or several three-dimensional) projection planes.

For practical applications of the analysis of the interrelationships of the parameters of individual sections of the cable car, it is important to be able to obtain projections of the phase trajectory of space with all variable parameters y and (Fig. 9).

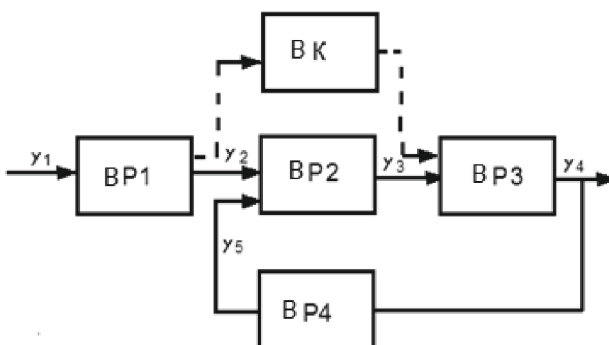


Fig. 11. Element block – diagrams of the control system of the suspended cable car

Analysis of the general picture of phase trajectory projections allows establishing such projections on both two-dimensional and three-dimensional projection planes, which have the largest variable component. Identifying problematic links in the control system allows you to install BK correction blocks (see Fig. 11), which make it possible to eliminate or reduce unwanted fluctuations. It is also possible to turn off the system element during the occurrence of a non-stationary process.

The second practically significant application of multidimensional phase spaces is the determination of the maximum values of parameters at a certain moment in time [13–15]. The problem boils down to determining the compromise extremum of the phase trajectory of the multidimensional phase space and can be solved by the algorithmic means developed. Note that k is a one-dimensional curved line ($k = 1$) or a hypersurface of a multidimensional phase space. The directional hyperplane, parallel to which the tangent hyperplane is drawn to the 1-manifold, i.e., the phase trajectory, or the hypersurface

of the phase space, is given by the equation in segments on the axes:

$$\frac{x_1}{\lambda_1} + \dots + \frac{x_i}{\lambda_i} + \dots + \frac{x_n}{\lambda_n} = 1, \quad (27)$$

where λ_i – the optimization weights.

In practical use for determining the maximum values of cableway parameters, it is important that the optimization criteria, i.e. the coefficients, are equivalent. Then the hyperplane (27) will be equally inclined to all axes of the multidimensional phase space.

6 Conclusions

The ultimate goal of this research is to improve geometric methods for process modelling, including the use of multiparameter (>3) geometric systems. This will also involve advanced data analysis techniques that have recently developed by authors, such as computer modeling using phase trajectory methods that help to analyze the dimensions of a complex system: this helps to inform the relationship between these parameters, which may not be independent. This can inform an understanding of the apparently unpredictable behaviour of such systems.

References

1. Made, R.I., et al.: Experimental characterization and modeling of the mechanical properties of Cu–Cu thermocompression bonds for three-dimensional integrated circuits. *Acta Mater.* **60**(2), 578–587 (2012)
2. Holt, J., Perry, S.A., Brownsword, M.: Model - Based Requirements Engineering
3. Ljaskovska, S., Martyn, Y., Malets, I., Prydatko, O.: Information technology of process modeling in the multiparameter systems. *IEEE Second Int. Conf. Data Stream Mining Process. (DSMP)* **2018**, 177–182 (2018)
4. Institution of Engineering and Technology, London, United Kingdom, p. 333 (2012)
5. Xioing, L., Ming, Y.: Simultaneous curve registration and clustering for functional data. *Comput. Stat. Data Anal.* **53**(4), 1361–1376 (2019)
6. Rama, A., Lekyasri, N., Rajani, K.: PID control design for second order systems. *IJEM* **9**(4), 45–56 (2019)
7. Yegul, M.F., Erenay, F.S., Striepea, S., Yavuza, M.: Improving configuration of complex production lines via simulation-based optimization. *Comput. Ind. Eng.* (109), 295–312 (2017)
8. Karimui, R.Y.: A new approach to measure the fractal dimension of a trajectory in the high-dimensional phase space. *Solitons Fractals* **151**, 111–239 (2021)
9. Zarei, M., et al.: Employing phase trajectory length concept as performance index in linear power oscillation damping controllers. *IEPE* **98**, 442–454 (2018)
10. Ali, S., Xie, Y.: The impact of industry 4.0 implementation on organizational behavior and corporate culture: the case of Pakistan’s retail industry. *IJEM*, 10(6), 20–31 (2020)
11. Peng, H., Zhu, Q.: Approximate evaluation of average downtime under an integrated approach of opportunistic maintenance for multi-component systems. *Comput. Ind. Eng.* **109**, 335–346 (2017)
12. Zarei, M., et al.: Oscillation damping of nonlinear control systems based on the phase trajectory length concept: an experimental case study on a cable-driven parallel robot. *Mech. Mach. Theory* **126**, 377–396 (2018)

13. Kalrath, J., Rebennack, S.: Cutting ellipses from area-minimizing rectangles. *J. Global Optim.* **59**(2–3), 405–437 (2014)
14. Savelyev, A.V., Stepanyan, I.V.: Spiral flows at the cardiovascular system as the experimental base of new cardiac-gadgets design. *IJEM* **8**(6), 1–12 (2018)
15. Giudici, P., Figini, S.: *Applied Data Mining for Business and Industry*. Wiley, p. 260 (2009)
16. Liaskovska, S., Izonin, I., Martyn, Y.: Investigation of anomalous situations in the machine-building industry using phase trajectories method. *ISEM* **463**, 49–59 (2021)



On Mikhailov Stability Conditions for a Class of Integer- and Commensurate Fractional-Order Discrete-Time Systems

Rafał Stanisławski^(✉) and Marek Rydel

Department of Electrical, Control and Computer Engineering,
Opole University of Technology, ul. Prószkowska 76, 45-758 Opole, Poland
{r.stanislawski,m.rydel}@po.edu.pl

Abstract. The Mikhailov stability condition is a classical stability test for continuous-time systems, similar to the well-known Nyquist method. However, in contrast to the Nyquist criterion, the Mikhailov stability tests do not reach considerable research attention. This paper introduces an extension of Mikhailov stability condition for two classes of a discrete-time system in terms of linear time-invariant system, and fractional-order one based on ‘forward-shifted’ Grünwald-Letnikov difference. Simulation experiments confirm the usefulness of the Mikhailov methods for stability analysis of discrete-time systems.

1 Introduction

The Mikhailov stability criterion is a classical graphical-oriented stability analysis method for dynamical systems. The method is based on the angle changes of a curve in the frequency domain of the system characteristic polynomial. The criterion was introduced by Mikhailov in 1938 for time-invariant continuous-time systems [1]. But, a few years later, analogical stability tests were independently proposed by Leonard [2] in 1944, and Cremer [3] in 1947. Therefore, the Mikhailov stability criterion is also called the Leonard or the Cremer-Leonard.

In contrast to the similar, Nyquist stability criterion, the Mikhailov test does not reach considerable research attention. Even though, this criterion is mentioned in several books related to systems and controls (see, e.g. [4]), but through many years this method does not reach any modifications to other classes of dynamical systems. Many years later, in Ref. [5] was introduced the modified Mikhailov stability criterion for a class of continuous, time-delayed systems. Recently, the research interest of Mikhailov stability criterion has significantly increasing in the context of fractional-order systems. The first paper in this area was devoted to continuous, time-delayed fractional-order systems [6]. Other modifications of Mikhailov tests were developed for continuous-time commensurate and incommensurate fractional-order systems [7,8]. The modification of Mikhailov stability criterion for specific, discrete-time nabla-based

fractional-order systems is presented in Ref. [9]. In the paper, we propose two Mikhailov stability conditions for two various classes of discrete-time systems. Firstly there is the proposed stability condition for a linear-time-invariant (integer-order) discrete-time system. Secondly, we propose stability results for a class of fractional-order discrete-time system with ‘forward-shifted’ Grünwald-Letnikov difference.

The paper is organized as follows. After a short introduction to the Mikhailov stability criterion in Sect. 1, Sect. 2 presents the original Mikhailov stability condition designed for continuous-time systems. On this basis, Sect. 3 introduces the Mikhailov stability criterion for the two mentioned discrete-time systems, which are the paper’s main results. Stability analysis examples of Sect. 4 present the effectiveness of the introduced test, and the conclusion of Sect. 5 finalizes the paper.

2 Mikhailov Stability Criterion

Mikhailov stability test is based on the curve resulting from the characteristic polynomial of the system in the complex plane. Therefore we consider a linear time-invariant continuous-time system with characteristic polynomial

$$p(s) = a_n s^n + \dots + a_1 s + a_0, \quad a_n > 0 \quad (1)$$

where s is Laplace transform operator. The Mikhailov condition defines that the system is stable if and only if the plot $p(j\omega)$, $\omega : 0 \rightarrow +\infty$ in frequency-domain satisfies the following two conditions [4]:

- a) $p(j0) = a_0 > 0$, i.e. the plot starts on the positive real axis,
- b) $\Delta \arg p(j\omega) \Big|_0^{+\infty} = \arg p(j\infty) - \arg p(j0) = \frac{n\pi}{2}$, i.e. as ω increases, the plot of $p(j\omega)$ encircles the origin in a counterclockwise direction and its phase goes to $n\frac{\pi}{2}$ for $\omega \rightarrow +\infty$.

As it is presented above, the stability test is based on the changes of the argument of the so-called Mikhailov curve $p(j\omega)$ for $\omega : 0 \rightarrow +\infty$. Presentation of Mikhailov curve for the system $p(s) = s^3 + 0.7s^2 + 0.64s + 0.058$ is demonstrated in Fig. 1.

It can be seen in Fig. 1 that $\arg p(j\omega)$ increases from 0 to quadrant III of the complex plane. Finally, $\Delta \arg p(j\omega) = \frac{3\pi}{2}$, therefore according to Mikhailov stability criterion, the system is asymptotically stable.

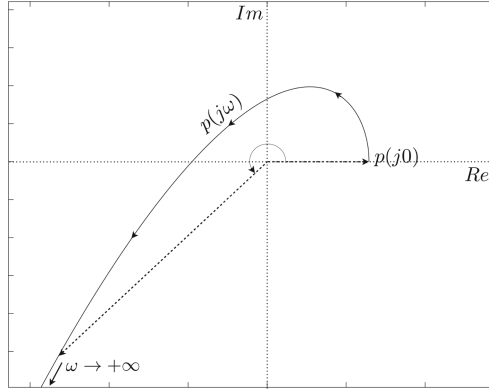


Fig. 1. Visualization of Mikhailov curve.

3 Main Result

3.1 Integer-Order Case

This paper introduces the Mikhailov stability criterion for linear time-invariant (LTI) discrete-time systems. As in the continuous-time case, the stability test is based on the characteristic polynomial of the system. Therefore, consider LTI discrete-time system with characteristic polynomial as follows

$$p(z) = a_n z^n + \dots + a_1 z + a_0 \tag{2}$$

where z is \mathcal{Z} -transform operator.

For the system with characteristic polynomial (2), we propose a modified Mikhailov stability condition for LTI discrete-time systems as follows.

Theorem 1. *Consider LTI discrete-time system with characteristic polynomial as in Eq. (2). Then the system is asymptotically stable if and only if*

- $p(e^{j\varphi}) \neq 0$, $\varphi = [0, \pi]$ and
- $\Delta \arg p(e^{j\varphi}) \Big|_0^\pi = n\pi$

where $p(e^{j\varphi}) : 0 \rightarrow \pi$ is so-called Mikhailov curve in the complex \mathcal{Z} -plane.

Proof. The characteristic polynomial of Eq. (2) can be presented in form

$$p(z) = a_n (z - \lambda_1) \dots (z - \lambda_n) = a_n \prod_{i=1}^n p_i(z) \tag{3}$$

where $p_i(z) = z - \lambda_i$, $i = 1, \dots, n$, and $\lambda_i \in \mathbb{C}$ are poles of the system. Consider various cases of location of poles of the system in the complex plane:

- a) if $\lambda_i \in \mathbb{R}$ and $|\lambda_i| < 1$ then $\Delta \arg p_i(e^{j\varphi}) \Big|_0^\pi = \arg p_i(e^{j\pi}) - \arg p_i(e^{j0}) = \pi$;

- b) if $\lambda_i \in \mathcal{R}$ and $|\lambda_i| > 1$ then $\Delta \arg p_i(e^{j\varphi})|_0^\pi = \arg p_i(e^{j\pi}) - \arg p_i(e^{j0}) = 0$;
 c) if $\lambda_i, \lambda_k \in \mathcal{C}$ with $\lambda_k = \lambda_i^*$, and $|\lambda_i| = |\lambda_k| < 1$ then $\Delta \arg[p_i(e^{j\varphi})p_k(e^{j\varphi})]|_0^\pi = \Delta \arg p_i(e^{j\varphi})|_0^\pi + \Delta \arg p_k(e^{j\varphi})|_0^\pi = 2\pi$;
 d) if $\lambda_i, \lambda_k \in \mathcal{C}$ with $\lambda_k = \lambda_i^*$, and $|\lambda_i| = |\lambda_k| > 1$ then $\Delta \arg[p_i(e^{j\varphi})p_k(e^{j\varphi})]|_0^\pi = \Delta \arg p_i(e^{j\varphi})|_0^\pi + \Delta \arg p_k(e^{j\varphi})|_0^\pi = 0$;
 e) if $|\lambda_i| = 1$ then $\exists \varphi_l: p_i(e^{j\varphi_l}) = p(e^{j\varphi_l}) = 0$;

Visualization of changes of $\Delta \arg p_i(e^{j\varphi})$ for stable and unstable poles is demonstrated in Fig. 2. On the basis on Eq. (3) we can easily show that

$$\Delta \arg p(e^{j\varphi})|_0^\pi = \sum_{i=1}^n \Delta \arg p_i(e^{j\varphi})|_0^\pi \quad (4)$$

Now, firstly, assume that we have $\exists l: p(e^{j\varphi_l}) = 0$, then $\exists i \in \{1, \dots, n\}: |\lambda_i| = 1$ (see condition e)), therefore the system is not asymptotically stable.

Secondly, assume that $|\lambda_i| \neq 1 \forall i \in \{1, \dots, n\}$ and the system have n_s stable poles and n_u unstable ones ($n = n_s + n_u$). Then on the basis on conditions a)–d) we have

$$\Delta \arg p(e^{j\varphi})|_0^\pi = \pi n_s \quad (5)$$

Taking into account that the system is asymptotically stable if and only if $n_u = 0$ and $n_s = n$, on the basis on Eq. (5) we immediately arrive at Theorem 1. \square

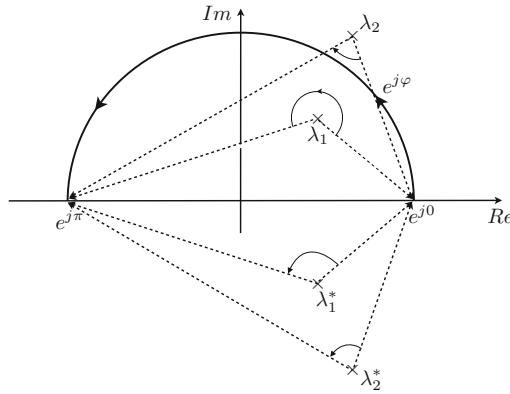


Fig. 2. Visualization of argument changes of $p_i(e^{j\varphi})$ for stable (λ_1) and unstable (λ_2) poles.

Note that the Mikhailov stability criterion for discrete-time systems is similar to continuous-time one. The main difference is that the condition for argument changes for the discrete-time case is two times larger than for continuous-time one. These results form the fact that the discrete-time stability area is closed

in the unit circle instead of the left-hand side of the complex plane. Also, note that, since $p(e^{j\pi}) < \infty$, the stability analysis for the discrete-time systems using the Mikhailov method is simpler than in continuous-time case. The usefulness of Mikhailov stability criterion for LTI discrete-time systems are presented in the next section.

3.2 Fractional-Order Case

Consider discrete-time commensurate fractional-order system presented in state-space form as

$$\begin{aligned}\Delta^\alpha x(t+1) &= Ax(t) + Bu(t) \\ y(t) &= Cx(t)\end{aligned}\tag{6}$$

where $t = 0, 1, 2, \dots$ is the discrete time, $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^{n_u}$, $y(t) \in \mathbb{R}^{n_y}$ are the state, input and output vectors, respectively, and $A \in \mathfrak{X}^{n \times n}$, $B \in \mathfrak{X}^{n \times n_u}$, $C \in \mathfrak{X}^{n_y \times n}$ are the system parameter matrices, respectively. The fractional-order difference $\Delta^\alpha x(t)$ is described by the use of Grünwald-Letnikov definition

$$\Delta^\alpha x(t+1) = \sum_{j=0}^{t+1} (-1)^j \binom{\alpha}{j} x(t+1-j)\tag{7}$$

In the similar way to the integer-order system, in the fractional-order case the stability analysis is based on characteristic pseudo-polynomial. It is well known, that the characteristic pseudo-polynomial for the system of Eq. (6) can be presented in the following form

$$\begin{aligned}p(z) &= a_n w^{\alpha n}(z) + a_{n-1} w^{\alpha(n-1)}(z) + \dots + a_0 \\ &= a_n (w^\alpha(z) - \lambda_1^f) (w^\alpha(z) - \lambda_2^f) \dots (w^\alpha(z) - \lambda_n^f) \\ &= a_n \prod_{j=1}^n p_j(z)\end{aligned}\tag{8}$$

where $w^\alpha(z) = z(1-z^{-1})^\alpha$, $p_j(z) = (w^\alpha(z) - \lambda_j^f)$, $j = 1, \dots, n$, and $\lambda_j^f \in \mathbb{C}$, $j = 1, \dots, n$, are the zeros of the characteristic pseudo-polynomial, which can be called pseudo-poles or, as in Refs. [10, 11], f -poles of the system (6).

For the system with characteristic polynomial (8), we can propose a modified Mikhailov stability condition for considered fractional-order discrete-time systems as follows.

Theorem 2. *Consider the discrete-time fractional-order system with characteristic pseudo-polynomial as in Eq. (8). Then the system is asymptotically stable if and only if*

- $p(e^{j\varphi}) \neq 0$, $\varphi = [0, \pi]$ and
- $\Delta \arg p(e^{i\varphi}) \Big|_0^\pi = n\pi$

where $p(e^{j\varphi}) : 0 \rightarrow \pi$ is so-called Mikhailov curve in the complex \mathcal{Z} -plane.

Proof. The characteristic pseudo-polynomial of Eq. (8) is described by elements $p_i(z) = w^\alpha(z) - \lambda_i$, $i = 1, \dots, n$ with $\lambda_i \in \mathbb{C}$ being f -poles of the system. Note that the system is asymptotically stable if and only if all f -poles there are inside the stability area (see Ref. [10]). Stability area \mathcal{S} for the system with fractional order $\alpha = 0.5$ is presented in Fig. 3. Now, consider various cases of the location of f -poles of the system in the complex plane:

- a) if $\lambda_i \in \mathcal{R}$ and $\lambda_i \in \mathcal{S}$ then $\Delta \arg p_i(e^{j\varphi}) \Big|_0^\pi = \arg p_i(e^{j\pi}) - \arg p_i(e^{j0}) = \pi$;
- b) if $\lambda_i \in \mathcal{R}$ and $\lambda_i \notin \mathcal{S}$ then $\Delta \arg p_i(e^{j\varphi}) \Big|_0^\pi = \arg p_i(e^{j\pi}) - \arg p_i(e^{j0}) = 0$;
- c) if $\lambda_i, \lambda_k \in \mathbb{C}$ with $\lambda_k = \lambda_i^*$, and $\lambda_{i,k} \in \mathcal{S}$ then $\Delta \arg [p_i(e^{j\varphi}) p_k(e^{j\varphi})] \Big|_0^\pi = \Delta \arg p_i(e^{j\varphi}) \Big|_0^\pi + \Delta \arg p_k(e^{j\varphi}) \Big|_0^\pi = 2\pi$;
- d) if $\lambda_i, \lambda_k \in \mathbb{C}$ with $\lambda_k = \lambda_i^*$, and $\lambda_{i,k} \notin \mathcal{S}$ then $\Delta \arg [p_i(e^{j\varphi}) p_k(e^{j\varphi})] \Big|_0^\pi = \Delta \arg p_i(e^{j\varphi}) \Big|_0^\pi + \Delta \arg p_k(e^{j\varphi}) \Big|_0^\pi = 0$;
- e) if $\exists \varphi_l: \lambda_i = w^\alpha(e^{j\varphi_l})$ then $p_i(e^{j\varphi_l}) = p(e^{j\varphi_l}) = 0$;

Visualization of changes of $\Delta \arg p_i(e^{j\varphi})$ for stable and unstable f -poles is demonstrated in Fig. 3. On the basis on characteristic pseudo-polynomial, on the same way as in integer-order case, we can easily show that

$$\Delta \arg p(e^{j\varphi}) \Big|_0^\pi = \sum_{i=1}^n \Delta \arg p_i(e^{j\varphi}) \Big|_0^\pi \quad (9)$$

Now, firstly, assume that we have $\exists l: p(e^{j\varphi_l}) = 0$, then $\exists i \in \{1, \dots, n\}: \lambda_i = w^\alpha(e^{j\varphi_l})$ (see condition e)), therefore the system is not asymptotically stable. Secondly, assume that $\lambda_i \neq w^\alpha(e^{j\varphi})$ and the system have n_s stable f -poles and n_u unstable ones ($n = n_s + n_u$). Then on the basis on conditions a)–d) we have

$$\Delta \arg p(e^{j\varphi}) \Big|_0^\pi = \pi n_s \quad (10)$$

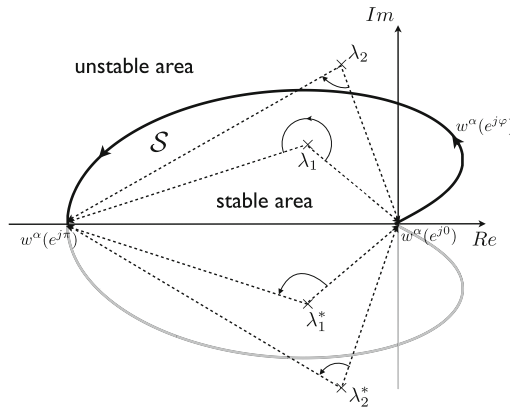


Fig. 3. Visualization of argument changes of $p_i(e^{j\varphi})$ for stable (λ_1) and unstable (λ_2) f -poles.

Taking into account that the system is asymptotically stable if and only if $n_u = 0$ and $n_s = n$, on the basis on Eq. (10) we immediately arrive at Theorem 2. \square

Remark 1. It is important to note, that the Mikhailov stability criterion is the same both for integer-order and fractional-order discrete-time systems with 'forward shifted' difference defined in Eq. (7),

Remark 2. It is also worth noticing, that definition of the fractional-order system of Eq. (6) is different from those presented in Ref. [9] and leads to different stability results based on the Mikhailov criterion.

4 Examples

Example 1. Consider a discrete-time LTI system with the characteristic polynomial

$$p(z) = 2.31z^3 - 2.35z^2 + z - \xi \quad (11)$$

with two different values of ξ , e.g. $\xi = 0.8$ and 1.05 , respectively. The poles of the system with both values of ξ are presented in Table 1.

Table 1 shows that the system is stable for $\xi = 0.8$, and unstable for $\xi = 1.05$. The Mikhailov curve for the considered system is presented in Fig. 4.

Table 1. Poles of the system for both values of ξ

ξ	$\lambda_{1,2}$	λ_3	Conclusion
0.8	$0.8336 \pm 0.6905i$	0.6828	stable
1.05	$0.6115 \pm 0.7468i$	1.127	unstable

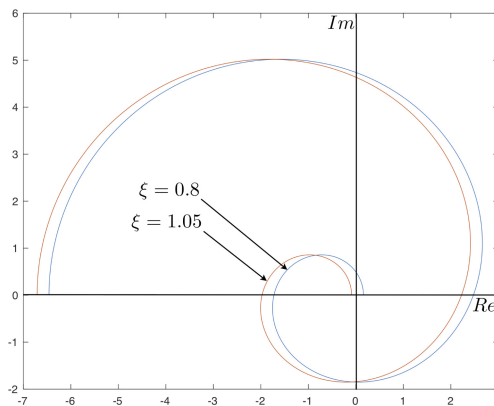


Fig. 4. Mikhailov curve for various values of ξ (Example 1).

We can see in Fig. 4 that for $\xi = 0.8$ we have $\Delta \arg p(e^{j\varphi}) = 3\pi$ (system is stable), but for $\xi = 1.05$ we have $\Delta \arg p(e^{j\varphi}) = 2\pi$ (system is unstable). Therefore, the Mikhailov stability test of Theorem 1 confirms stability of the system.

Example 2. Consider a discrete-time LTI system with the characteristic polynomial

$$p(z) = z^4 - 2z^3 + \xi z^2 - 1.13z + 0.3074 \tag{12}$$

with various values of $\xi = 1.95$ and 2.3 , respectively. The poles of the system with both values of ξ are presented in Table 2.

Table 2. Poles of the system for both values of ξ

ξ	1.95	2.3
$\lambda_{1,2}$	$0.3000 \pm 0.7000i$	$0.6579 \pm 0.8331i$
$\lambda_{3,4}$	$0.7000 \pm 0.2000i$	$0.3421 \pm 0.3946i$
Conclusion	stable	unstable

The Mikhailov curve for the system with two various values of ξ is presented in Fig. 5

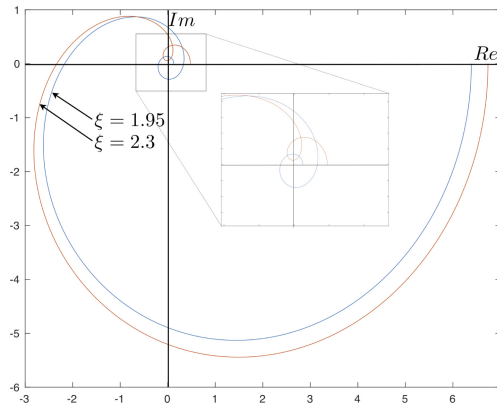


Fig. 5. Mikhailov curve for various values of ξ (Example 2).

Table 2 shows that the system is stable for $\xi = 1.95$, and it is unstable for $\xi = 2.3$. The Mikhailov curve of Fig. 5 shows that for $\xi = 1.95$ we have $\Delta \arg p(e^{j\varphi}) = 4\pi$ so the system is asymptotically stable. In contrast, for $\xi = 2.3$ we have

$\Delta \arg p(e^{j\varphi}) = 2\pi$, therefore on the basis of Theorem 1, we can conclude that the system has two unstable poles and therefore it is unstable. The Mikhailov-based results in both cases fully correspond to the results obtained based on poles' location.

Example 3. Consider a discrete-time LTI system with the characteristic polynomial

$$p(z) = z^3 - 1.8z^2 + 1.32z - 0.2 \quad (13)$$

The zeros of characteristic polynomial are $\lambda_1 = -0.2$ and $\lambda_{2,3} = 0.8 \pm 0.6i$, so the system is not asymptotically stable ($|\lambda_{2,3}| = 1$). The Mikhailov curve is presented in Fig. 6.

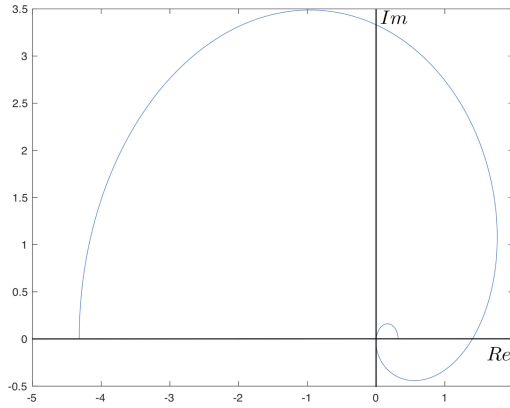


Fig. 6. Mikhailov curve for Example 3.

It can be seen in Fig. 6 that Mikhailov curve $p(e^{j\varphi})$, $\varphi : 0 \rightarrow \pi$, passes through the origin of the complex plane. Therefore taking into account Theorem 1, the system is not asymptotically stable. This, again, confirms the effectiveness of the Mikhailov stability criterion.

Example 4. Consider fractional-order discrete-time state space system of Eq. (6) with matrices

$$\left[\begin{array}{c|c} A & B \\ \hline C & D \end{array} \right] = \left[\begin{array}{c|c} 0 & -\xi & 1 \\ \hline 1 & 0 & 0 \\ 0 & 1 & 0 \end{array} \right]$$

and fractional order $\alpha = 0.5$ and $\xi = 0.95$ and 1.1 , respectively. The characteristic pseudo-polynomial is as $p(z) = w^1(z) + \xi$. The Mikhailov curves for two various values of ξ are presented in Fig. 7.

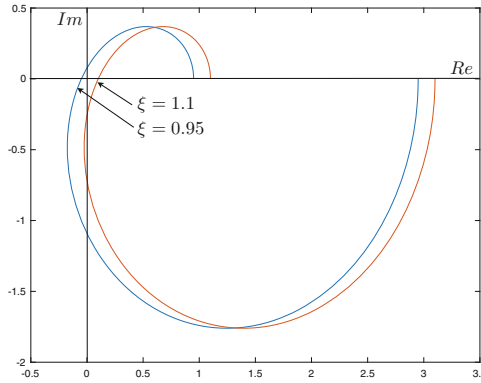


Fig. 7. Mikhailov curve for Example 4.

We can see from Fig. 7 that $\Delta \arg p(e^{j\varphi}) = 2\pi$ for $\xi = 0.95$ and the system is asymptotically stable. In contrast, for $\xi = 1.1$ we have $\Delta \arg p(e^{j\varphi}) = 0$ and the system is unstable. Accounting that pseudo-poles of the system are $\lambda_{1,2} = \pm 0.9747i$ for $\xi = 0.95$ and $\lambda_{1,2} = \pm 1.0488i$ for $\xi = 1.1$, this result can be easily confirmed by the analytical condition proposed in Theorem 4 of Ref. [11].

5 Conclusion

The paper has presented the implementation of the Mikhailov stability criterion for both, the LTI discrete-time system and a class of commensurate fractional-order systems. The main results of Theorems 1 and 2 developed the simple stability tests for the system based on argument changes of the so-called Mikhailov curve, which is based on (pseudo-)characteristic polynomial. Interestingly, the Mikhailov stability conditions for considered fractional-order systems are the same as for the LTI integer-order one.

References

1. Mikhailov, A.: Method of harmonic analysis in control theory. *Avtomatika i Telemekhanika* **3**, 27–81 (1938)
2. Leonhard, A.: Neues verfahren zur stabilitätsuntersuchung, *Archiv für Elektrotechnik* **38**(17–28) (1944)
3. Cremer, L.: Ein neues verfahren zur beurteilung der stabilität linearer regelungssysteme, *ZAMM* **167** (1947)
4. Ackermann, J.: *Robust Control. Systems with Uncertain Physical Parameters*. Springer, London (1993). <https://doi.org/10.1007/978-1-4471-3365-0>
5. Barker, L.K.: Mikhailov stability criterion for time-delayed systems, *NASA Technical Memorandum*, p. 78803 (1979)
6. Busłowicz, M.: Stability of linear continuous-time fractional order systems with delays of the retarded type. *Bull. Polish Acad. Sci. Tech. Sci.* **56**(4), 319–324 (2008)

7. Mendiola-Fuentes, J., Melchor-Aguilar, D.: Modification of Mikhailov stability criterion for fractional commensurate order systems. *J. Franklin Institute* **355** (2018)
8. Stanisławski, R.: Modified Mikhailov stability criterion for continuous-time non-commensurate fractional-order systems. *J. Franklin Inst.* **359**, 1677–1688 (2022)
9. Stanisławski, R., Latawiec, K.J.: A modified Mikhailov stability criterion for a class of discrete-time noncommensurate fractional-order systems. *Commun. Nonlinear Sci. Numer. Simul.* **96**, 105697 (2021)
10. Stanisławski, R., Latawiec, K.J.: Stability analysis for discrete-time fractional-order LTI state-space systems. Part I: new necessary and sufficient conditions for asymptotic stability. *Bull. Polish Acad. Sci. Tech. Sci.* **62**, 353–361 (2014)
11. Stanisławski, R., Latawiec, K.J.: Stability analysis for discrete-time fractional-order LTI state-space systems. Part II: new stability criterion for FD-based systems. *Bull. Polish Acad. Sci. Tech. Sci.* **62**, 362–370 (2014)



Verification of a Building Simulator in Real Experiments

Karol Jabłoński¹ and Dariusz Bismor²(✉)

¹ Atest Gaz, Gliwice, Poland
kkf.jablonski@gmail.com

² Silesian University of Technology, Gliwice, Poland
Dariusz.Bismor@polsl.pl

Abstract. The increase of energy prices has a substantial influence on the need of development of efficient control algorithms for the HVAC systems. However, testing of such algorithms in real buildings is usually expensive due to long time constants and unpredictable and unrepeatable weather conditions. Therefore, building simulation software plays an important role in the process. In this paper, the design of a building HVAC system simulator developed within the “Stiliger” project is presented. The simulator operates based on the data including all the necessary building construction details. The paper compares the results of the simulations with the measurements on a real, existing building, which was modeled. The comparison shows good agreement of the indoor temperatures in the simulation and in the real building.

1 Introduction

With the increase of energy prices, energy efficient control algorithms for heating, ventilation and air conditioning (HVAC) systems became an important part of a solution to the energy savings problems. However, development and testing of such algorithms in real buildings is expensive and hard to perform, given the very slow dynamic of physical buildings. Therefore, building simulators are used to speed up development and testing phases of such research [1]. One of such simulators was created by the authors within the “Stiliger” project, using Matlab/Simulink® software [2, 7]. The simulator allows to simulate the most important parameters of a building from the point of view of a HVAC control algorithm, including temperatures of walls, room temperatures, and carbon dioxide concentration. To be as realistic as possible, all building partitions are simulated as multi-layered, according to their actual construction. The simulator has a flexible structure, allowing to simulate any building, provided the building construction project is available.

In the research described in this chapter, an existing, real physical building project was implemented in the simulator, and the simulations were compared with real measurements of various building temperatures. The building was a single-story single-family house located in southern Poland. It was equipped

with a floor heating system, charged by a heat pump. The heat pump control system was modified by application of the Stiliger controller — a microcontroller dedicated to HVAC systems developed within the “Stiliger” project. This allowed to apply exactly the same control algorithm as in the simulation software. The building was equipped with a number of temperature sensors, and the system was able to record the temperatures. This in turn allowed to compare the results obtained using the simulator with behavior of the building.

The paper is organized as follows. The mathematical models used in the development of the simulator, including the heat exchanger models, are presented in Sect. 2. The developed simulation software is described in Sect. 3. The building where the tests were conducted is described in Sect. 4. The comparison of the simulation results and the real building measurements is presented in Sect. 5. Finally, the paper is concluded in Sect. 6.

2 Mathematical Models

2.1 Modeling of Building Elements

For the purposes of the simulation, the building was decomposed into individual rooms, and each room into individual walls and the air inside. These elements are modeled using ordinary differential equations with lumped parameters, which are appropriate to simulate buildings behavior [5], even if gray box models are also frequently used with good results [3, 9].

Changes of temperature of air in the i -th room are described by the equation resulting from the energy balance:

$$\frac{dT_i}{dt} = \frac{1}{C_p} [Q_p + Q_v + Q_e + Q_g + Q_d + Q_w]; \quad (1)$$

where C_p is the heat capacity of air in the room, Q_p is the energy supplied by heating systems (if present), Q_v is the energy supplied by the air from a ventilation system, Q_e is the energy emitted by people staying in the room, Q_g is the energy transferred through windows, and Q_w is the energy radiated from the walls.

Changes of temperature of the j -th layer of n -th wall in i -th room (floor and ceiling are also considered as wall) are modeled by the following equation:

$$\frac{dT_{inj}^w}{dt} = \frac{A_{ni}}{C_{ni}} \left[U_{inj}^{iw} (T_{in(j+1)}^w - T_{inj}^w) + U_{inj}^{ow} (T_{in(j-1)}^w - T_{inj}^w) + \frac{Q}{A_{ni}} \right]; \quad (2)$$

where A_{ni} is the area of the wall, C_{ni} is its heat capacity, U_{inj}^{iw} , U_{inj}^{ow} are the inner and outer thermal conductivities of the wall, and Q is an additional heat delivered to the wall, for example from sun radiation or floor heating.

2.2 Heat Exchanger Modeling

Many heating systems used in residential buildings are based on heat exchangers with a liquid heating medium. In proposed solution the equations were derived from the basic laws of thermodynamics [6], but there are also other approaches, such as approximation used by Skruch [8]. The heat flux supplied to the ambient by this type of exchanger is:

$$\frac{dQ}{dt} = UA(T_v - T_{avg}); \quad (3)$$

where:

A is area of exchange, U is thermal exchange coefficient, T_v is ambient temperature, and T_{avg} is average temperature of heating medium.

A simplified model was adopted as the average temperature of the medium in the installation T_{avg} , which is the arithmetic mean of the temperature of the medium at the inlet and outlet of the installation:

$$T_{avg} = \frac{T_{in} + T_{out}}{2}. \quad (4)$$

The heat flux emitted by the exchanger can also be expressed as:

$$\frac{dQ}{dt} = c_w \rho q (T_{in} - T_{out}); \quad (5)$$

where: c_w is specific heat of the heating medium, ρ is density of the heating medium, q is volume flow of the heating medium, T_v is ambient temperature, and T_{avg} is average temperature of the heating medium.

If the measurement of the temperature of the heating medium at the exchanger outlet (T_{out}) is not available, you can calculate this value by equating (3) to (5). We then get:

$$T_{out} = \frac{(\frac{c_w \rho q}{UA} - 0.5)T_{in} + T_v}{\frac{c_w \rho q}{UA} + 0.5}. \quad (6)$$

3 The Developed Simulation Software

During the development, various technologies and programming languages were used for better organization and optimization of the simulator individual elements. This approach facilitates creation and subsequent modifications of the algorithm. It also allows to optimize code execution performance. Data preparation using real measurements was carried out using Linux shell scripts, the main simulation component was designed as a Simulink schematic diagram, and mathematical models of elements were implemented as Matlab s-functions in the C language. The simulation is launched by a Matlab m-file script, which allows to control various simulation parameters. All source files and data files was ordered in a hierarchical way.

3.1 Structure of the Simulator

The subsystems nested and ordered into libraries were used in the design of the Simulink model. Their organization and relations are presented in Fig 1. The master system includes the building simulator subsystem, data inputs from the workspace, and the subsystems for ventilation, heating (radiators, floor heating), and air conditioning. This layer is also prepared for implementation of control algorithms.

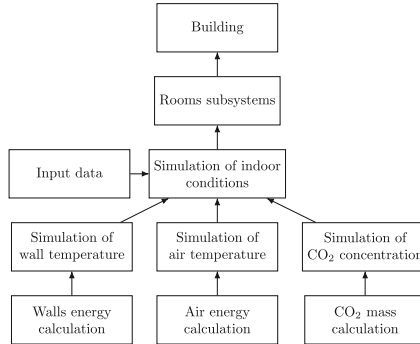


Fig. 1. The structure of the simulator.

The building subsystem accepts the following inputs (one element of the vector corresponds to one room):

- vector of air flow values from ventilation;
- vector of supply air temperatures;
- vector of occupant numbers;
- vector of heating source powers;
- vector of additional powers for walls (Q in Eq. (2));
- outdoor air temperature;
- soil temperature;
- vector of solar radiation on vertical surfaces in four directions.

The subsystem returns the following outputs:

- vector of carbon dioxide concentrations;
- vector of indoor air temperatures;
- bus of vectors of wall temperatures.

The building subsystem contains subsystems of individual rooms, which have outputs analogous to the outputs of the entire building. The input is a bus signal, consisting of all the building inputs and outputs.

Each room subsystem consists of two elements. The first is a selector that extracts parameters corresponding to the particular room from the input bus,

and a module that calculates the amount of solar irradiation entering the room through its windows. The second consists of s-functions to solve the respective differential equations for calculation of the air temperature, wall temperatures, and the concentration of carbon dioxide.

3.2 Data Organization

It turned out to be quite a challenge to construct a simulation template that would allow the modeling of buildings with different geometries. For this purpose, a special system of arrays, structures and objects as well as indexing of rooms and walls has been developed [2]. Such organization allows room subsystems to easily extract data to be used in simulation. All input data should be also placed in format readable for Simulink. Figure 2 presents the scheme of data organization.

To simulate the energy balance, knowledge of conductances and thermal capacities of construction elements and air is required. The structures containing these values can be filled in manually, but to simplify the process, functions that accept the basic construction parameters of a room (density, thickness, thermal conductivity, specific heat of wall layers) have been prepared, and then the thermal capacities and conductivities are calculated.

4 The Test Building Structure

The test building is a single-family, one-story building with an undeveloped attic, located near geographical coordinates 50.2, 18.7. Its plan is shown in Fig. 3.

The building consists of 10 rooms on the ground floor: vestibule (1), toilets (2), office (3), bathroom (4), bedrooms (5,6), communication space (7), living room (8), technical room (9) and kitchen (10). In the simulation, for the purposes of data organization, the attic (11), outdoor space (12) and the the ground (13) were also distinguished. The interior of the building is strongly “open”, which was a problem during the simulation: the kitchen, living room and hall are practically one room.

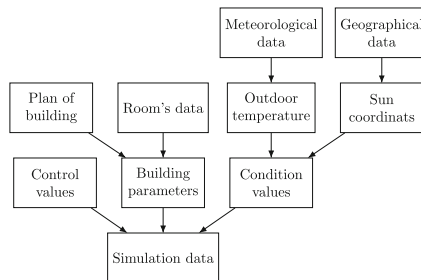


Fig. 2. Data for the simulation.



Fig. 4. The internal structure of the building.

Table 1. Floor structure.

No.	Layer	d	λ	R	ρ	c_w	C
		m	W/m·K	m ² K/W	kg/m ³	J/(kg·K)	J/m ² · K
1	floor panels	0.08	0.220	0.090	800	2510	160640
2	concrete (B20)	0.15	1.300	0.115	2000	840	252000
3	polystyrene (FS20)	0.16	0.036	4.444	20	1460	4672
4	concrete (B20)	0.10	1.300	0.070	2000	840	168000
5	sand	0.20	0.400	0.500	1650	840	277200

Table 2. External walls structure.

No.	Layer	d	λ	R	ρ	c_w	C
		m	W/m·K	m ² K/W	kg/m ³	J/(kg·K)	J/m ² K
1	mineral plaster	0.01	0.820	0.012	1850	840	15540
2	polystyrene	0.12	0.038	3.157	13.5	1030	1668
3	expanded clay concrete	0.15	0.540	0.277	1200	840	151200
4	interior plaster and stoneware tiles	0.01	0.820	0.012	1850	840	15540

The external walls are made of prefabricated expanded clay concrete slabs. From the outside, the walls are insulated with polystyrene. Detailed parameters of external walls are presented in the Table 2.

The internal walls were also made of prefabricated elements with a thickness of 10 cm, and their parameters are presented in the Table 3.

The ceiling above the residential part of the building is made of plasterboard and insulated with a layer of mineral wool. Detailed parameters of the ceiling are shown in the Table 4.

Table 3. Internal walls structure.

No.	Layer	d	λ	R	ρ	c_v	C
		m	W/m·K	m ² K/W	kg/m ³	J/(kg·K)	J/m ² K
1	mineral plaster	0.01	0.820	0.012	1850	840	15540
2	expanded clay concrete	0.10	0.540	0.185	1200	840	100800
3	mineral plaster	0.01	0.820	0.012	1850	840	15540

Table 4. Ceiling structure.

No.	Layer	d	λ	R	ρ	c_v	C
		m	W/m·K	m ² K/W	kg/m ³	J/(kg·K)	J/m ² K
1	drywall	0.01	0.230	0.043	1000	1000	10000
2	mineral wool	0.10	0.043	2.325	60	750	4500
3	wood	0.01	0.230	0.043	1000	1000	10000

5 Experiment Results

5.1 Measurements

The experiment in the test building began on April 2, 2020 and lasted for three weeks. During this time the house was heated by the underfloor heating charged by a heat pump. The system was controlled by the dedicated microcontroller device prepared within the “Stiliger” project and programmed with algorithm presented by Bismor et al. [2].

During the tests, the temperatures at the following points were measured:

- supply temperature of the floor heating circuit, at the outlet from the heat pump,
- return temperature of the floor heating circuit (at the inlet to the heat pump),
- utility hot water temperature (on the tank jacket),
- outdoor temperature (north elevation of the building),
- indoor temperature in the living room (no 8),
- coolant supply and return temperatures,
- three temperatures on the compressor gas circuit.

Sample measurement results for the period from April 17 to April 23 are shown in Fig. 5.

The figure shows that the range of measured temperatures starts from negative temperatures (outside temperature) and ends at a temperature exceeding 90°C (temperature on the gas circuit of the heat pump, just after the compressor). The waveforms reveal a number of periodic phenomena, e.g. daily changes in the outdoor temperature, clearly visible heating cycles, periods when utility water was heated, etc.

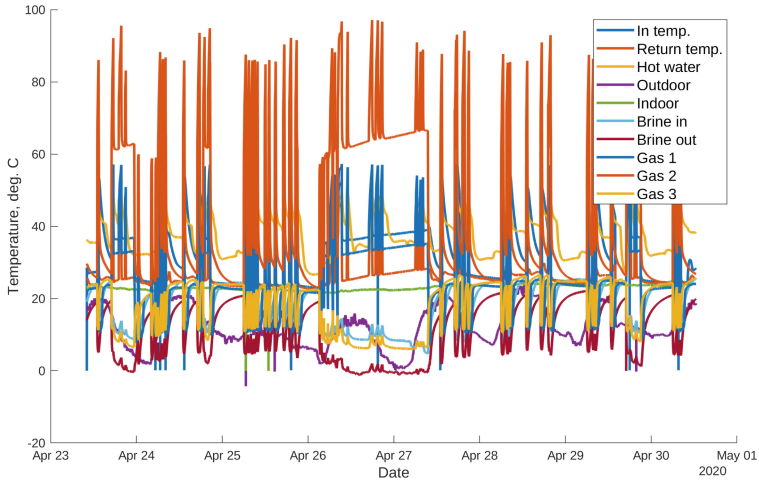


Fig. 5. Temperature measurements.

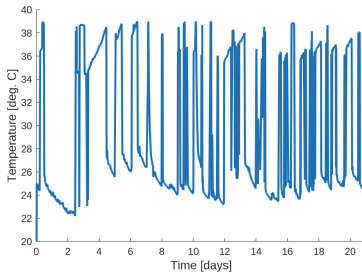


Fig. 6. Temperature of medium in the floor heating during the experiments.

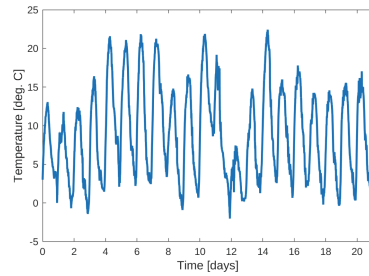


Fig. 7. Outdoor temperature during the experiments.

5.2 Simulation and Results

In the simulations presented below, the measured temperature at the outlet of the heat pump and the measured outdoor temperature were used to feed the model described in Sect. 2 and parametrized like in Sect. 4. Both the temperatures are presented in Figs. 6 and 7. The simulator also used information based on observations of the occurrence of one of the five most common insolation scenarios on a given day:

1. cloudy all day;
2. clouds in the morning, sunshine in the afternoon;
3. morning 1/4 cloudy, around noon 1/2, then 1/4;
4. around noon 1/4 cloudy, the rest sunny;
5. all day sunny.

The output of the model was a vector of temperature measurements in specific rooms at consecutive time instants. The summary of results is presented in

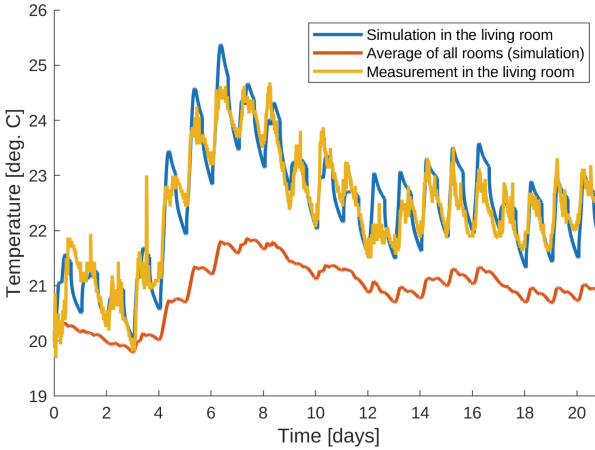


Fig. 8. Temperature inside the building.

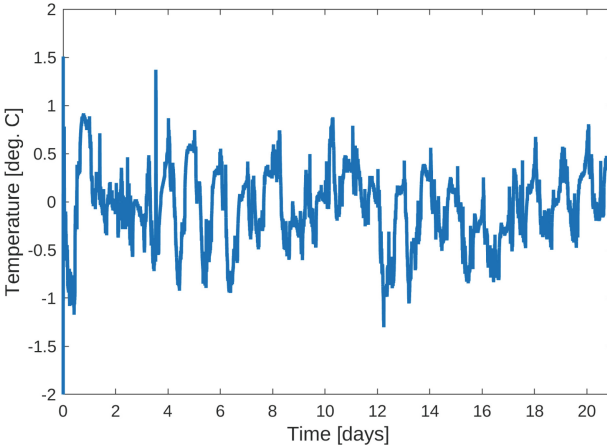


Fig. 9. Difference between measurements and model.

Fig. 8, where the blue curve presents result of the simulation in the living room and the yellow one presents the actual measured temperature. The figure reveals high fluctuations of the indoor temperature, which were caused by large area of windows. This in turn allows to appreciate the impact of solar radiation on the indoor temperature. The mean temperature of all rooms was also presented in the figure as a red curve, and, as it can be observed that the mean temperature does not show similar degree of variation.

Figure 6 presents temperature of underfloor heating during experiment and Fig. 7 shows outdoor temperature.

On order to quantify the performance of the simulation system, the difference between the measurements in the living room and the model output in the living

room was calculated and is presented in Fig. 9. For the majority of the simulation time (which was over 20 d), the absolute value of the difference between the two temperatures does not exceeds 1°C . The calculated mean absolute difference was only 0.3°C .

6 Conclusions

In this paper, the building simulator created in the Matlab/Simulink® was evaluated by comparison of the simulation results with the measurements in a real building. As the simulator uses the actual building design parameters, the model was expected to be in a good agreement with the measurements. The experiments confirmed this supposition, showing only a small differences, mainly due to high influence of the solar radiation through large areas of the south-facing windows.

Acknowledgment. The work described in this paper is within the project “Synergiczny system automatyki budynkowej zintegrowany z układami optymalizacji komfortu i klimatu w budynkach—SSAB”, which is co-sponsored by WND-RPSL.01.02.00-24-0853/17-001, Regional Operating Program for Silesian Voivodship for years 2014–2020.

References

1. Andersen, K.K., Madsen, H., Hansen, L.H.: Modelling the heat dynamics of a building using stochastic differential equations. *Energy Build.* **31**, 13–24 (2000)
2. Bismor, D., Jabłoński, K., Grychowski, T., Nas, S.: Hardware-in-the-loop simulations of a GPC-based controller in different types of buildings using Node-RED. In: Bartoszewicz, A., Kabziński, J., Kacprzyk, J. (eds.) *Advanced, Contemporary Control*. AISC, vol. 1196, pp. 1018–1029. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-50936-1_85
3. Ferracuti, F., et al.: Data-driven models for short-term thermal behaviour prediction in real buildings. *Appl. Energy* **204**, 1375–1387 (2017). <https://doi.org/10.1016/j.apenergy.2017.05.015>. <https://www.sciencedirect.com/science/article/pii/S0306261917305032>
4. Garbalińska, H., Bochenek, M.: Izolacyjność termiczna a akumulacyjność cieplna wybranych materiałów ściennych. *Czasopismo Techniczne. Architektura* **108**(2-A/2), 89–96 (2011)
5. Hudson, G., Underwood, C.: A simple building modelling procedure for MATLAB/SIMULINK. In: *Proceedings of the International Building Performance and Simulation Conference, Kyoto Japan*, vol. 2, pp. 777–783. Citeseer (1999)
6. Incropera, F.P., DeWitt, D.P., Bergman, T.L., Lavine, A.S., et al.: *Fundamentals of Heat and Mass Transfer*, vol. 6. Wiley, New York (1996)
7. Perera, D., Winkler, D., Skeie, N.O.: Multi-floor building heating models in MATLAB and Modelica environments. *Appl. Energy* **171**, 46–57 (2016). <https://doi.org/10.1016/j.apenergy.2016.02.143>

8. Skruch, P.: A thermal model of the building for the design of temperature control algorithms. *Automatyka/Automatics* **18**(1), 9–21 (2014)
9. Thilker, C.A., Bacher, P., Bergsteinsson, H.G., Junker, R.G., Cali, D., Madsen, H.: Non-linear grey-box modelling for heat dynamics of buildings. *Energy Build.* **252** (2021). www.scopus.com. Cited By :13



Aspects of Measurement Data Acquisition and Optimisation in the Energy Transformation of Industrial Facilities

Lukasz Korus and Andrzej Jabłoński^(✉)

Faculty of Information and Communication Technology, Department of Control Systems and Mechatronics, Wrocław University of Science and Technology, Wrocław, Poland

{lukasz.korus, andrzej.jablonski}@pwr.edu.pl

Abstract. This article deals with the problem of energy management, understood as the effort to minimise the energy consumption in the industrial facilities by continuous improvement approach. Due to the fact that multi-level Decentralised Control Systems (DCS), with the main focus on optimising energy consumption in small and medium enterprises are still rarely used, the continuous improvement approach based on Deming-like cycle using data collected by Data Acquisition and Presentation Systems from measuring devices is often used in practice. This paper presents a high-level process used in companies to drive energy management in an efficient way, and the main objectives defined for such an approach. It also provides more practical information on infrastructure configurations, IT components and applications of data collection and presentation systems.

1 Introduction

One of the challenges facing the world today is to take all possible measures to stop climate warming. Such a challenge implies a radical reduction in greenhouse gas emissions, and one of the ways to do this is to transform the world's energy system. Energy transformation is a multidimensional process, because it includes, among others, the issues of minimising energy consumption, eliminating hydrocarbon energy sources, introducing ecological, so-called green energy sources, creating energy storage facilities, but also developing new technologies/industrial equipment [5, 11]. The field of energy transformation includes the private sphere, the public sphere and the broadly understood industrial/business sphere. The introduction of energy transformation requires scientific research in many fields and the systematic implementation of its results for everyday use. It should be noted that the effective implementation of the energy transition is connected with the skilful combination of scientific methods of identification, simulation and modelling, optimisation, artificial intelligence with the issues of data acquisition and presentation, information and communication technologies (ICT) [2], technological and construction solutions, but also with the support

of activities in the field of economics, management, sociology and psychology. This article reduces the complex process of energy transformation to selected aspects of measurement data acquisition and the possibility of optimising electricity management in industrial facilities and technological processes [3].

2 Energy Management: Process and Goals

This section contains information about high level business goals that are defined for energy management (in particular for: data acquisition and processing) and general approach that might be used for continues improvement in energy management.

2.1 Continues Improvement Cycle

The Deming Cycle, or PDCA approach, was proposed by the quality control pioneer and describes a basic, iterative process used to solve problems or support continuous improvement. PDCA stands for Plan-Do-Check-Act and defines four steps of the approach. The first step - plan - involves investigating and analysing to fully understand the situation. Based on the information gathered, the plan is prepared, the desired state and results are defined. It is worth mentioning that the plan is prepared based on the knowledge acquired at a certain point in time and does not necessarily cover all aspects and details. This basically means that the knowledge and understanding of the situation is still limited. In the next step - do, all changes defined in the plan are implemented with one important assumption - they should be small scale, not forcing significant changes in the system or environment. It should be a small step in the defined direction with minimal disruption. The next step - check - is dedicated to in-depth analysis of the situation after implementation of the planned changes. The effects of the changes introduced must be monitored, described and understood. The results of the investigation are used in the next step - act, where the additional corrections are introduced based on the suggestions from the previous step, in order to introduce further improvements in the defined direction. This approach has been a foundation for agile techniques widely used in industry and IT [6]. The Deming Cycle can be easily modified or adopted [4] to create a simple flow that supports continuous improvement in efficient energy management (Fig. 1). In this case we have one improvement step instead of two as in the PDCA approach, but two important activities have been introduced: measure, prepare. The first describes the process of collecting information from sensors (e.g. energy, gas, flow, environmental meters, etc.), while the second is devoted to the preparation of all activities aimed at reducing energy consumption in the analysed object. In an ideal situation, this cycle can be implemented using a control system with a feedback loop, but in many cases it ends with significant changes in the structure of the object (e.g. replacement of machines, equipment, etc.). The last step in the cycle, called act, can be considered as the introduction of structural adjustments or actions of effectors that directly affect the object, as in control theory.

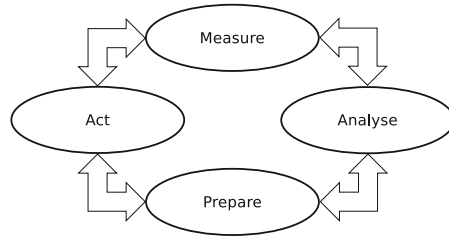


Fig. 1. Continues improvement approach in energy management.

2.2 Business Goals

In practice, the main objective of energy management, which is to reduce energy consumption in small and medium-sized manufacturing companies, is rarely achieved through the use of multi-level control systems. However, based on a good understanding of the distribution of energy consumers, some structural adjustments can be introduced. Due to this fact, the rest of the paper will focus more on the process of data collection and processing. From this point of view, the following high-level business objectives for the system can be specified:

- building knowledge about the current state of the object by gathering energy and environmental measurements (short-term behaviour, current data),
- building awareness of the object's behaviour in different circumstances (long-term behaviour, historical data),
- analysis of the dependencies between data sets in the power grid (data analysis),
- notifying end users about defined situations in the power grid (notifications),
- forecasting object's behaviour based on the historical data (data forecasting),
- possibility of direct impact on the object by manual control (manual control),
- optimisation of energy consumption by control systems (automatic control).

The high level business objectives listed above should be broken down into a list of functionalities and further into functional and non-functional requirements. There is no point in listing all the low-level requirements here, but as an example, the list of some non-functional requirements is given below. Non-functional requirements are frequently connected to the numbers that can be specified to describe the system:

- number of measurement points defined in the system,
- measurement cycle for each of the measurement points,
- measurement period for each of the measurement points,
- structure of data in the system, so the question if the data will be further consolidated in more complex structures.

Based on the non-functional requirements mentioned above, the architect of the system can take crucial decision about the architectural concept to secure acceptable performance of the system once it's implemented. For example: let's assume

that we have ten energy meters with only one channel each. Let's assume that the data will be read once a minute and store in data base. It is easy to calculate that having only ten devices, the number of data collected per day is 14400 and for the year, it is more than 5000000. Based on this simple calculation of amount of data in the system for ten meters only, one can easily understand that it is a crucial aspect of such type of systems impacting performance negatively, if it is not properly addressed. For example, using cloud environment might not be the best choice since it will be connected with growing, significant cost. In order to limit this issue, it is worth to consider edge computing and storing raw data only locally. Moreover, in order to prepare reporting functionality it is recommended to use some BI methods and tools. One of the interesting examples of the systems that can cover some of the needs mentioned above, would be Azure IoT Hub which might be easily integrate with BI report. Unfortunately, it will not solve all the technical and financial challenges [9].

3 System Architecture

Taking all the aspects mentioned in the previous sections into consideration, simple architecture of the data acquisition and presentation system can be proposed (Fig. 2). As you can see in the picture, there are two types of infrastructure or ways of collecting data from metering devices. One is based on standard Ethernet networks. The other is based on industrial wireless communication standards such as Long Range Wide Area Networks (LoRaWAN) or Bluetooth Low Energy (BLE). In practice, the situation is not as simple as shown in the figure, mainly because of the variety of communication protocols and standards available in the instruments. Some devices are equipped with the RS-485 communication standard and Modbus or M-Bus protocols, while others are much more advanced and use the Device Language Message Specification (DLMS) standard. DLMS is a global standard used to communicate with smart energy meters. Unfortunately, DLMS is not an open standard and it is necessary to purchase the specification to be able to implement it properly. In summary, the infrastructure part of the system is usually much more complicated and includes various electronic devices such as converters and hubs that translate information from one protocol to another or store information from many devices to increase redundancy in the system. More information about the modern infrastructure approach will be mentioned in the following sections of this paper. On top of the low-level communication layer, the Message Queuing Telemetry Transport (MQTT) protocol is often used to implement efficient communication based on the Publisher-Subscriber way of exchanging information. In practice, to ensure communication based on the MQTT protocol, an MQTT broker (e.g. Mosquitto) needs to be visible in the network and all measurement devices should have access to the broker and be able to communicate using MQTT, either directly or through the converters or hubs. Using MQTT is efficient, secure and convenient because it is a modern and popular standard and there are many additional IT tools that can simplify the presentation of data directly from the broker. Unfortunately,

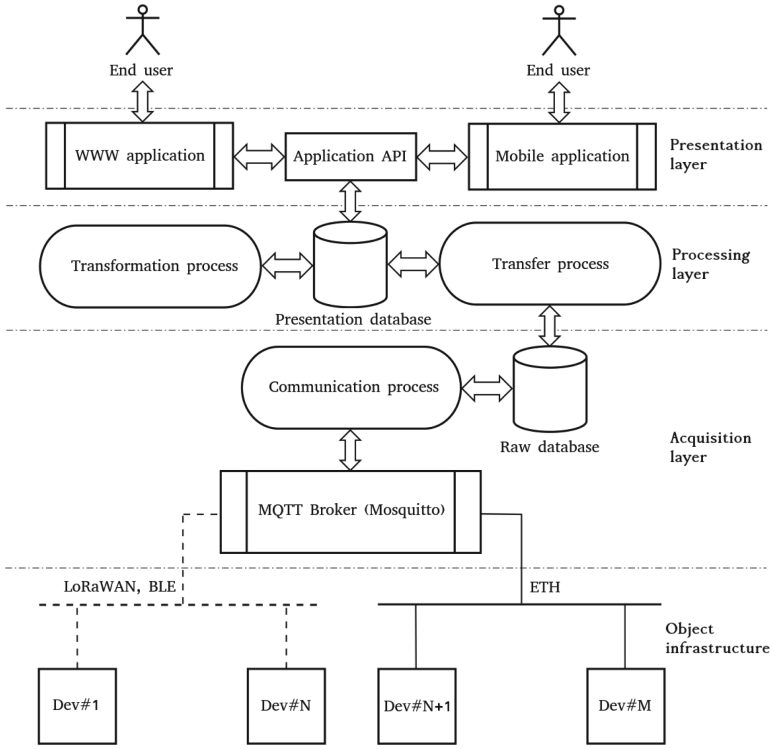


Fig. 2. Architecture of data acquisition and presentation system.

the data sent to the MQTT broker is stored for a limited period of time, and there is a need to move it to a more stable environment. One of the options is to transfer the data to the SQL database (e.g. MySQL) using the communication process. This database is used to store raw data for a limited period of time (e.g. a week). After this period, the data should be transferred to the archiving database, but this data will not be visible in the system through the user interface. The next step is to properly process the raw data in terms of values and frequency of measurements. This basically means that some of the data can be multiplied by a constant value to reflect the infrastructure setup, or some of the data can be cumulated or removed to keep a smaller amount of data in the presentation database. Together with these pre-processing activities, the data is transferred to the presentation database. This database contains information that will be presented to the end users of the system using different technologies (www, mobile). Usually, even if this pre-processing has taken place, it is very difficult to present all the data in a way that ensures acceptable performance. For example, it would be difficult to produce a report containing a year's worth of 15-minute energy profiles from a measurement point, because it would be about 34176 of measurements with time stamp and status. So, none of the known

technologies, especially www, will be able to present such amount of data in an efficient way (even if paging is implemented). For this reason, the data should be further processed using e.g. BI tools to prepare data sets for the reports or figures defined in the system. In addition, if there is a notification subsystem that notifies end users when the defined situation occurs, the data for this subsystem should be computed in the background so as not to affect the performance of the presentation layer. The last, but not least part of the system is the presentation layer, which consists of an application API and various types of presentation applications (e.g. based on www or mobile technologies). In practice, one of the modern Java Enterprise Edition technologies such as Spring Boot can be used to implement the back-end application API and React JS framework to implement the front-end part of the system.

4 Infrastructure

This section contains more detailed information about two simple, practical use cases [7] prepared to present real and working infrastructure configurations. The first is a simple data acquisition and presentation system created based on the industrial version of Raspberry PI and Modbus on RS-485 communication. The second is more advanced and based on RAK LoRaWAN devices. The first one can be used for simpler use cases with limited number of energy issues and simple application functionalities. The second is more scalable and can be used in large-scale systems that collect data from a large number of metering devices spread over a large area.

4.1 Use Case #1: Simple Hub with Modbus on RS-485 Communication

Figure 3 shows a simple data acquisition and presentation system based on an RPi4 computing module and communication implemented based on Modbus on RS-485. This is a basic infrastructure configuration, but it could be useful in most simple use cases where there is a need to collect information from a limited number of energy meters and present the data within a limited period of time. On the other hand, this type of infrastructure may be part of a larger system. In such cases, the RPi4 module is treated as a hub, collecting data from a local group of energy meters and ensuring redundancy of data and infrastructure in the system. The rest of the paper presents the more general and scalable architecture.

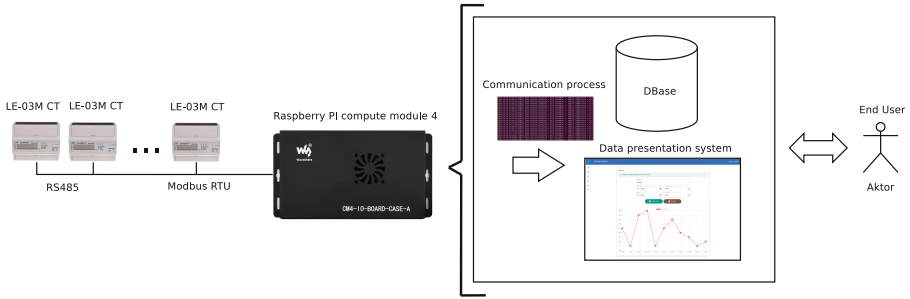


Fig. 3. Simple data acquisition and presentation system based on RPi4 compute module.

4.2 Use Case #2: LoRaWAN Based System

Figure 4 presents data acquisition and presentation system with LoRaWAN communication. This solution is much more scalable and flexible, also due to the fact that it uses low-power radio communication [8]. As can be seen in the Fig. 3, RAK's equipment was used in this configuration to secure the LoRaWAN communication. Starting from the left, there are RAK7431 converters that use Modbus RTU protocol on RS-485 bus to collect information from energy meters. RAK7431 as a master device is equipped with the functionality that supports communication with many slave devices one by one in the cycle. This basically ensures the possibility of reading information from many energy meters or other devices connected to the RS-485 bus. Once the information is read from a particular device, there are at least two ways to send it further: either the information is tunneled through the LoRaWAN protocol and sent to the LoRa Gateway (RAK7289), or it is tunneled through MQTT and sent to the Gateway (RAK7289), which in this case is treated as an MQTT broker. In the first option, the information can be sent further to the external broker using the MQTT protocol over WiFi or LTE. In the second option, since the gateway

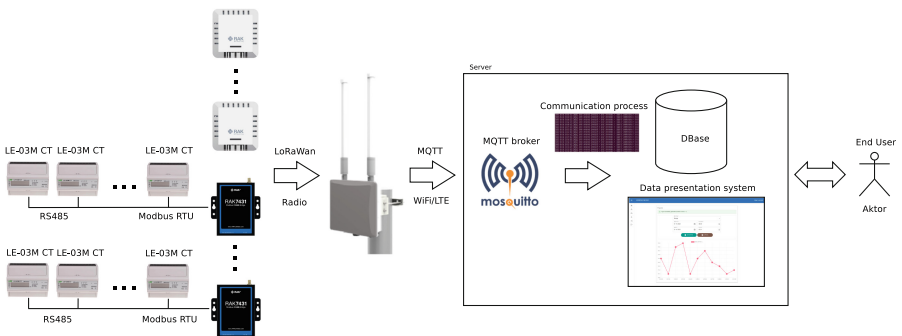


Fig. 4. Data acquisition and presentation system based on LoRaWAN.

(RAK7289) plays the role of MQTT broker, the presentation systems should be equipped with subscriber functionality to be able to receive data from the broker. Both options are equally scalable and can be used in different situations to meet different requirements, but the ultimate goal is to get the measurement data stored in the database. In the next step, the raw data collected from the measurement devices can be further processed and finally visualised in different forms on the presentation layer/application. It can be a web, mobile or desktop application, but it's important to ensure scalability, performance and access for multiple users. Another challenge is how to present such a huge amount of data using the chosen user interface technology. It's obvious that all the data cannot be visible at once, as this would cause some performance problems. So, it is necessary to think about additional algorithms that would ensure aggregation of data with acceptable speed and quality.

5 General Structures

Based on the previous considerations, two general approaches can be defined: a classic one based on the master-slave configuration often used in the industry, and a more modern one based on the publisher-subscriber approach using IoT devices and the MQTT protocol. While there are many significant differences in configuration and data flow between these two approaches, one challenge is common to both configurations. In the complex systems with many measurement points and high frequency of reading data from them, there is a risk that the cost of the cloud environment will grow in an uncontrolled way and the performance will decrease significantly. For this reason, it may be necessary to consider an edge computing type approach. In practice, this means that data should first be collected on the local machine in the factory, all data processing must be done locally and only aggregated data can be sent to the cloud environment. On the one hand, this ensures sufficient performance and much lower operating costs of the system, and on the other hand, it gives the possibility to present only high-level, aggregated data. When it comes to the differences between these two types of infrastructure and communication, the first is based on a pull approach, which means that the slave devices only send data at the request of the master. The second is based on a push approach, which in practice means that each of the devices can send data simultaneously and the broker has to manage this communication efficiently. There is one more thing that is visible in both solutions and it increases the redundancy in the systems. There are two servers and two databases, as mentioned in the previous section. The first server, installed locally in the factory, is responsible for maintaining uninterrupted communication with the gauges and storing the raw data in the database. The second, which could be in the cloud environment, stores data ready for presentation. It's also worth mentioning that on the first server we store configuration data describing the network of measuring devices, while the second is used to store business configuration data. Moreover, there are two separate networks in which these machines operate: the first operates within a separate measurement/industrial network,

while the second operates within the company's internal network. Both types of configuration ensure acceptable performance, security and good maintainability and scalability of the entire solution (Fig. 5).

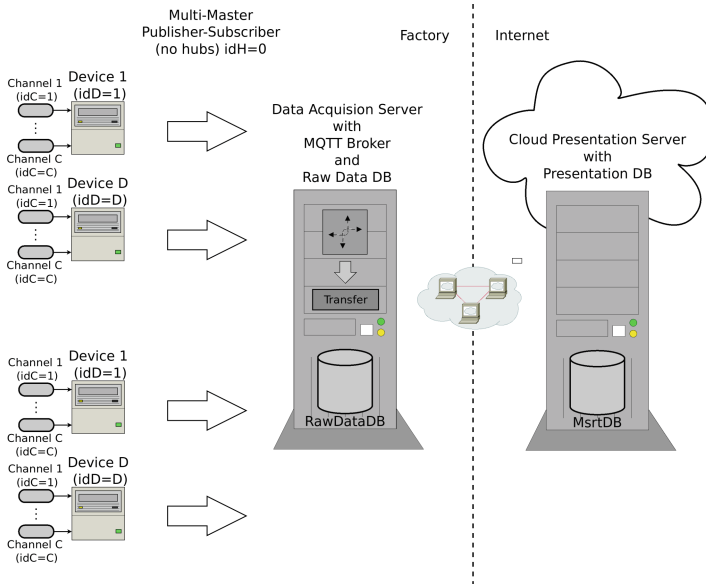


Fig. 5. Data acquisition and presentation system based on Publisher-Subscriber architecture.

6 Summary

Energy management or energy optimisation methods in small and medium-sized industrial companies are still rarely based on optimal control algorithms with objective functions focused on minimising the total energy in the production process. This has not been the case in the past, because the income from production was much higher than the energy costs. At the moment, the situation is becoming much more difficult, taking into account the fact that energy costs are many times higher than they were a few months ago. It is not difficult to foresee that more attention will be paid to optimal control algorithms that minimise energy consumption. However, before that happens, the easiest way to deal with this issue now is to collect information from sensors that measure energy consumption at different points in the factory and use a continuous improvement approach to eliminate or limit energy leakage. Modern data collection and presentation systems are a great help in this regard, and can ultimately lead to a better understanding of the distribution and balance of energy consumption in the factory. All statistical methods of data analysis can be used to draw

conclusions based on the measurements (time series) [1]. With this understanding, one can either make some changes in the infrastructure of the factory or use some effectors to control the highest energy consumers in the factory. The recommendation for further development is to use IoT systems with MQTT protocol to gather data from energy meters and use it for further processing in order to build models, calculate predications and control energy consumption. The disadvantage of this approach is separation from DCS and this needs to be addressed.

References

1. Korus, L., Piorek, M.: Compound method of time series classification. *Nonlinear Analysis: Modelling Control* **20**(4), 545–560 (2019). <https://doi.org/10.15388/NA.2015.4.6>
2. Yang, C., Vyatkin, V.: Design and validation of distributed control with decentralized intelligence in process industries: a survey. In: *IEEE International Conference on Industrial Informatics (INDIN)*, pp. 1395–1400 (2008)
3. Geddes, K.O., Czapor, S.R., Labahn, G.: *Energy Efficiency in the Process Industries. A User-guide to Sustainable Energy Efficiency*, Emerson (2014)
4. Prashar, A.: Adopting PDCA (Plan-Do-Check-Act) cycle for energy optimization in energy-intensive SMEs. *J. Clean. Prod.* **145**, 277–293 (2017)
5. Frasnier, J., Morea, F., Krenn, C., Uson, J.A., Tomasi, F.: Energy efficiency in small and medium enterprises: Lessons learned from 280 energy audits across Europe. *J. Clean. Prod.* **142**(4), 1650–1660 (2017)
6. Parrish, K., Whelton, M.: *LEAN operations: An Energy Management Perspective*, Proceedings IGLC-21, pp. 865–874, Fortaleza, Brazil (2013)
7. Ferencz, K., Domokos, J.: IoT sensor data acquisition and storage system using raspberry pi and apache cassandra. In: *2018 International IEEE Conference and Workshop in Óbuda on Electrical and Power Engineering (CANDO-EPE)*, pp. 143–146, Budapest, Hungary (2018). <https://doi.org/10.1109/CANDO-EPE.2018.8601139>
8. Haxhibeqiri, J., De Poorter, E., Moerman, I., Hoebeke, J.: A survey of LoRaWAN for IoT: from technology to application. *Sensors* **18**(11), 1424–8220 (2018)
9. Forsström S., Jennehag U.: A performance and cost evaluation of combining OPC-UA and Microsoft Azure IoT Hub into an industrial Internet-of-Things system, *Global Internet of Things Summit (GIoTS)*, Geneva, Switzerland (2017), pp. 1–6. <https://doi.org/10.1109/GIOTS.2017.8016265>.
10. Xuan, S., DoHyeun, K.: Performance analysis of IoT services based on clouds for context data acquisition. In: Lee, R. (ed.) *BCD 2018*. *SCI*, vol. 786, pp. 81–92. Springer, Cham (2019). https://doi.org/10.1007/978-3-319-96803-2_7
11. Gorzyński, J.: *Efektywnosc energetyczna w dzialalnosci gospodarczej*. Wydawnictwo Naukowe PWN, Warszawa (2023)



Continuous-Time Dynamic Model Identification Using Binary-Valued Observations of Input and Output Signals

Jarosław Figwer^(✉)

Department of Measurements and Control, Silesian University of Technology,
Akademicka 16, 44-100 Gliwice, Poland

Jaroslaw.Figwer@polsl.pl

Abstract. In the paper an approach to the identification of linear continuous-time dynamic systems based on acquired binary-valued observations of their input and output signals is presented. In the presented approach the binary-valued observations are interpreted as disturbed measurements of the corresponding original continuous-time input and output signals and system models are estimated using classical linear dynamic system estimation methods. The properties of such the approach are illustrated with three simulation examples. In these examples, frequency responses are identified using the empirical transfer function estimator, parametric models are recovered from the frequency responses identified, and finally, parametric models are identified directly from the binary-valued observations acquired.

1 Introduction

In modern digital world there are gathered big data in which exist pieces of information possible to interpret as binary-valued observations of continuous-time signals. They can be used for model identification. Considerations presented in the paper are devoted to parametric as well as nonparametric continuous-time model identification of linear dynamic systems based on binary-valued observations of the corresponding continuous-time input and output signals. Extensive literature research resulted in no publications discussing such the identification problem. An attentive reader can find a few publications in which the identification of discrete-time parametric models of linear dynamic systems, that are only FIR filters, based on the binary-valued observations of input and output signals is discussed — see e.g. [4, 6] and references included therein. Additionally, it is worth to emphasise that the identification problem formulated in these publications is artificial one and the discussed methods of solving it have very restricted applications.

The paper consists of two parts. The first part is devoted to a formulation of the linear continuous-time dynamic system model identification problem in which the binary-valued observations are used and to a proposition of an approach to the corresponding model identification. In the second part, three simulation examples

are presented to illustrate the properties of the presented approach. They show how to identify the frequency response using the empirical transfer function estimator, how to recover the corresponding parametric model from the frequency response identified, and finally, how to identify the parametric model directly from the binary-valued observations acquired. The considerations presented below are an addendum to the considerations presented in the monograph [3].

2 Identification Problem Formulation

Let us assume that the plant to be identified is a stable, rational and time-invariant continuous-time single-input single-output linear dynamic system described by the following differential equation:

$$\sum_{\nu=0}^{dA} a_{\nu} \frac{d^{\nu} y(t)}{dt^{\nu}} = \sum_{\nu=0}^{dB} b_{\nu} \frac{d^{\nu} u(t - T_d)}{dt^{\nu}}, \quad (1)$$

where:

- t ($t \in \mathcal{R}$) denotes the continuous time,
- $u(t)$ is the continuous-time input signal,
- $y(t)$ is the continuous-time output signal,
- dA and dB ($dA \geq dB$) being the non-negative integers are the structure numbers of the plant,
- $a_0, a_1, \dots, a_{dA}, b_0, b_1, \dots, b_{dB}$ are the coefficients of the above differential equation,
- T_d is the time-delay.

In the above differential equation, the zero-order differentiation returns the continuous-time input or output signal, respectively. It is assumed that the continuous-time input signal $u(t)$ is a wide-sense stationary Gaussian random process having the zero-mean value.

The purpose of the identification is to determine an estimate of the frequency response of the system (1), and/or estimates of the coefficients $a_0, a_1, \dots, a_{dA}, b_0, b_1, \dots, b_{dB}$ and the time-delay T_d based on N binary-valued observations of the continuous-time signals $u(t)$ and $y(t)$ acquired with the sampling interval T . This means that the result of an identification experiment are the following sets of the binary-valued observations:

$$\{v(0), v(1), \dots, v(N-1)\}, \quad (2)$$

$$\{\gamma(0), \gamma(1), \dots, \gamma(N-1)\}, \quad (3)$$

where $v(i) = \mathcal{F}(u(iT))$ and $\gamma(i) = \mathcal{F}(y(iT))$ for $i = 0, 1, \dots, N-1$. The $\mathcal{F}(\cdot)$ is a static nonlinear transformation defined as:

$$\mathcal{F}(x) = \begin{cases} up & ; x \geq 0 \\ lo & ; x < 0 \end{cases}, \quad (4)$$

where up and lo are real-valued numbers satisfying the relation $up > lo$.

It is worth noticing that the above sets of the binary-valued observations represent infinitely many different continuous-time linear dynamic systems having the same poles, zeros and time-delays but differencing with the steady-state gains.

Though the disturbances at the plant output were omitted in Eq. (1), this the acquired binary-valued observations of the continuous-time input and output signals are the original continuous-time signal values distorted by the transformation $\mathcal{F}(\cdot)$. This transformation of the continuous-time input or output signal may be represented as a sum of the two components: one being the signal under transformation, and other being a random process that is uncorrelated with the signal under transformation. It implies that the above-formulated identification problem is a continuous-time linear dynamic system identification problem in which the binary-valued observations of the continuous-time input and output signals are interpreted as disturbed measurements of the original continuous-time input and output signals. The level of these disturbances depends on the choice of up and lo values. It can be expressed as a signal-to-noise ratio defined as the ratio of the variance of the signal under transformation to the corresponding variance of the disturbances. For example, a calculation of the such defined signal-to-noise ratios for 10000 realisations of a discrete-time Gaussian white noise with the variance equal to 1.0000 and the number of samples equal to 131072 resulted in the mean values equal to 2.4738 for $up = -lo = 1.0000$ and equal to 0.0555 for $up = -lo = 5.0000$. It follows from this interpretation of the transformation $\mathcal{F}(\cdot)$ that the above-formulated identification problem is an identification problem in which the corresponding linear dynamic model is identified based on measurements being samples of the disturbed continuous-time input and output signals. It is worth to emphasise that the appearance at the plant output, in Eq. (1), of random disturbances with the zero-mean value will no influences on the presented interpretation of the above-formulated identification problem.

It is also worth to mention that the case in which the continuous-time input signal $u(t)$ is precisely measured, no processed by the transformation $\mathcal{F}(\cdot)$, is not discussed in the paper. The reason is: it is a problem of the linear dynamic subsystem model estimation of Wiener system that is widely discussed in many publications.

3 Simulation Examples

Below, the identification of the continuous-time linear dynamic plant models using the binary-valued observations was illustrated with three simulation examples. In these examples frequency responses were identified using the empirical transfer function estimator, parametric models were recovered from the frequency responses identified, and finally, parametric models were identified directly from the binary-valued observations acquired. Continuous-time linear dynamic plants were simulated using Runge-Kutta-Fehlberg method [5]. The

corresponding binary-valued observations of the continuous-time input and output signals were acquired with the sampling interval T equal to 0.1000 [s] using the transformation \mathcal{F} with the parameter $up = -lo = 5.0000$. It implies that in the simulation examples the values of the signal-to-noise ratio implied by the transformation $\mathcal{F}(\cdot)$ were extremely very low. As the continuous-time input signal $u(t)$ realisations of the band-limited to the range $[0, 5]$ [Hz] white continuous-time multisine random signal with the variance equal to 1.0000 were applied [1–3]. These realisations were synthesised and simulated assuming that their periodograms (calculated using samples of the continuous-time signal) are equal to periodograms of the realisations of the discrete-time Gaussian white noise with the variance equal to 1.0000 for all multiplicities of the frequency bin in the frequency range $[0, 5]$ [Hz]. The above-mentioned models were identified using ideas described in [3].

3.1 Frequency Response Identification

In the first simulation example a continuous-time linear dynamic plant with the following transfer function:

$$K(s) = \frac{1.5000}{1.2000s^2 + 1.0000s + 1.0000} e^{-0.3300s} \quad (5)$$

was simulated resulting in N binary-valued observations of the continuous-time signals $u(t)$ and $y(t)$. These observations were divided into nonoverlapping data segments of the length $M = 1024$ each. For the each data segment the frequency response was estimated using the empirical transfer function estimator, i.e. for all relative frequencies Ωm ($\Omega = \frac{2\pi}{M}$ and $m = 0, 1, \dots, M - 1$) the estimate of the frequency response was calculated as the ratio of finite discrete Fourier transforms of the corresponding binary-valued observations of the continuous-time signals $y(t)$ and $u(t)$. The resulting frequency response estimate for the relative frequencies Ωm ($m = 0, 1, \dots, M - 1$) was obtained by averaging all the results of estimation for this relative frequency obtained for the consecutive data segments. In Fig. 1 the results of the frequency response identification for 100 realisations of the band-limited to the range $[0, 5]$ [Hz] white continuous-time multisine random signal and $N = 102400$ are presented. In Fig. 2 the corresponding results obtained for $N = 1024000$ are shown. In Fig. 3 the mean value of these results is compared with the pattern. A similar comparison of the normalised frequency responses, having the steady-state gain equal to 1.0000, is in Fig 4.

3.2 Parametric Model Recovery from Frequency Response

The frequency response identified is a good starting point to recover the corresponding parameters of the continuous-time transfer function. The mentioned parameters may be obtained by minimizing the objective function being a sum of the squared distances between the frequency response identified and the frequency response calculated from the continuous-time transfer function for the

angular frequencies $\frac{\Omega}{T}m$ ($m = 0, 1, \dots, \frac{M}{2} - 1$). Using the frequency responses identified in the previous simulation example for $N = 1024000$ binary-valued observations the corresponding parameters (one coefficient of the numerator, three coefficients of the denominator and the time-delay) of the continuous-time transfer function were recovered. In Fig. 5 the frequency response estimates calculated using 100 sets of continuous-time transfer function parameters recovered are presented. In Fig. 6 the mean value of these results is compared with the pattern. Comparison of the corresponding normalised frequency responses, having the steady-state gain equal to 1.0000, is in Fig 7.

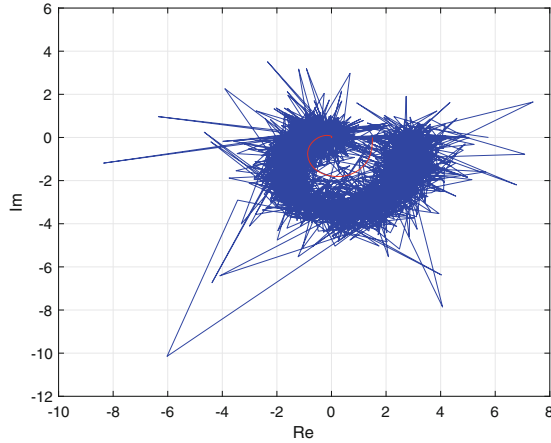


Fig. 1. 100 frequency response estimates calculated using the empirical transfer function estimator (blue lines) and the pattern (red line) — $N = 102400$, $M = 1024$.

3.3 Parametric Model Identification

In the third simulation example a continuous-time linear dynamic plant with the following transfer function:

$$K(s) = \frac{2.5000s + 2.5000}{2.0000s^2 + 3.0000s + 1.0000} e^{-0.8880s} \quad (6)$$

was simulated resulting in $N = 262144$ binary-valued observations of the continuous-time signals $u(t)$ and $y(t)$. During model identification the binary-valued observations of the continuous-time output signal calculated from the corresponding model excited by the binary-valued observations of the continuous-time signal $u(t)$ were fitted to the acquired binary-valued observations of $y(t)$. During this fitting a sum of the squared differences between the calculated and the acquired binary-valued observations of the signal $y(t)$ was minimised. The

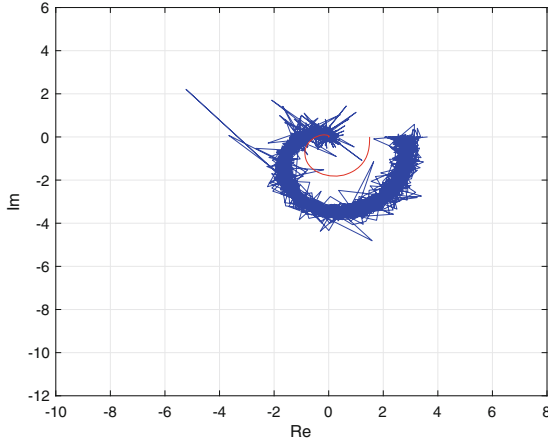


Fig. 2. 100 frequency response estimates calculated using the empirical transfer function estimator (blue lines) and the pattern (red line) — $N = 1024000$, $M = 1024$.

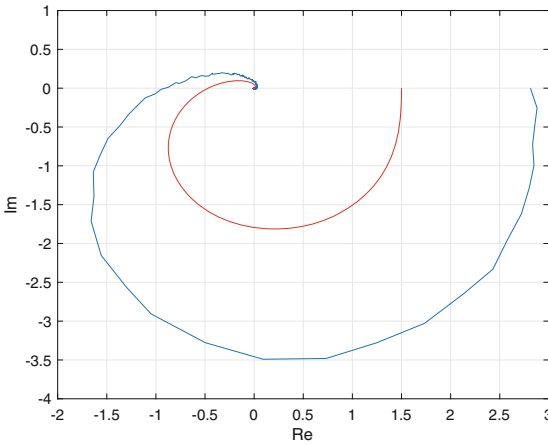


Fig. 3. Mean value of 100 frequency response estimates calculated using the empirical transfer function estimator (blue line) and the pattern (red line) — $N = 1024000$, $M = 1024$.

result of this minimisation were estimates of the parameters (two coefficients of the numerator, three coefficients of the denominator and the time-delay) of the above continuous-time transfer function. Using these parameters, the unit step response of the corresponding model was calculated. The sequence of operations consisting of acquiring the binary-valued observations, the parameter estimation and the unit step response calculation was repeated 100 times. In Fig. 8 100 unit step responses calculated based on the estimated parameters are presented.

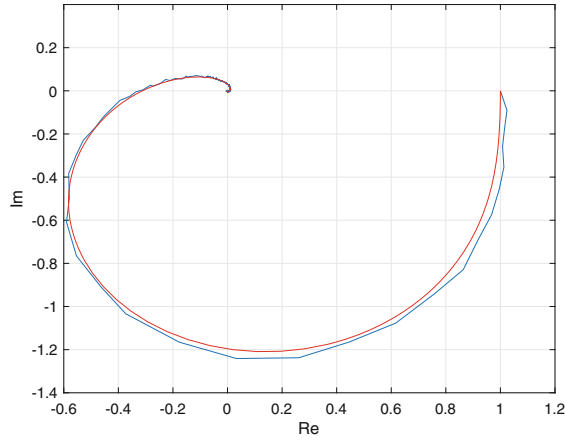


Fig. 4. Normalised mean value of 100 frequency response estimates calculated using the empirical transfer function estimator (blue line) and the normalised pattern (red line) — $N = 1024000$, $M = 1024$.

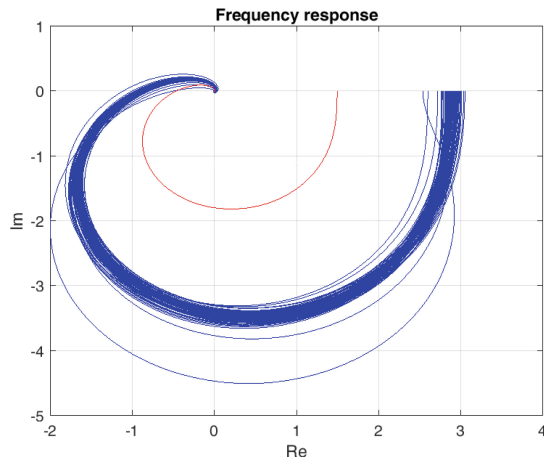


Fig. 5. 100 frequency response estimates calculated from the parametric models recovered (blue lines) and the pattern (red line) — $N = 1024000$, $M = 1024$.

The corresponding mean value and the pattern are in Fig. 9. In Fig. 10 a comparison of the corresponding normalised, having the steady-state gain equal to 1.0000, mean value of the calculated step responses with the normalised pattern is shown.

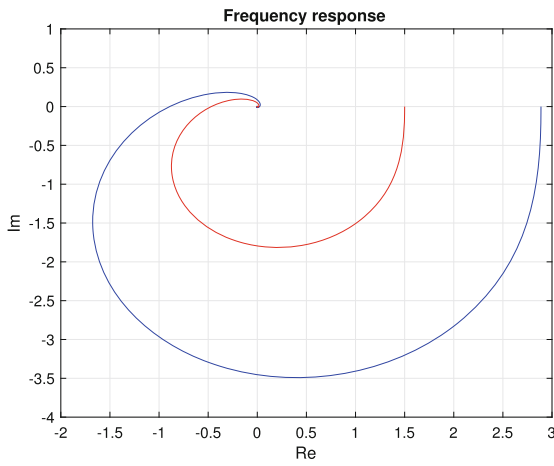


Fig. 6. Mean value of 100 frequency response estimates calculated from the parametric models recovered (blue line) and the pattern (red line) — $N = 1024000$, $M = 1024$.

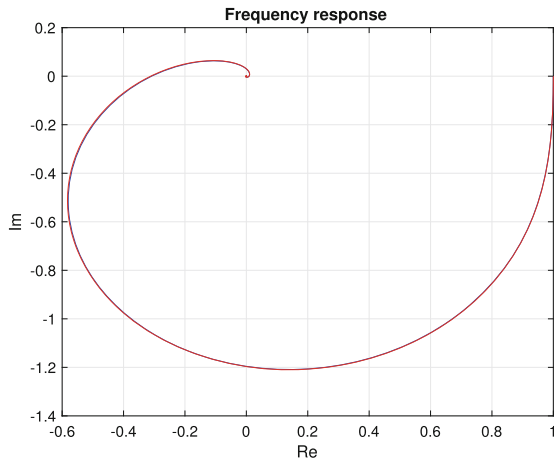


Fig. 7. Normalised mean value of 100 frequency response estimates calculated from the parametric models recovered (blue line) and the normalised pattern (red line) — $N = 1024000$, $M = 1024$.

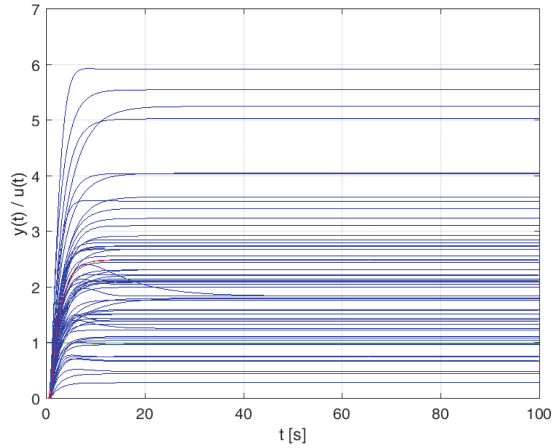


Fig. 8. 100 unit step responses calculated from the parametric models identified (blue lines), the pattern (red line) and the unit step excitation (green line) — $N = 262144$.

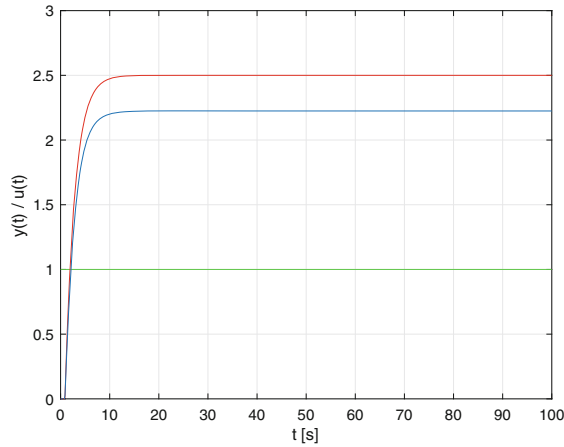


Fig. 9. Mean value of 100 unit step responses calculated from the parametric models identified (blue line), the pattern (red line) and the unit step excitation (green line) — $N = 262144$.

4 Summary

In the above-presented simulation examples obtained results of identification are reported without a discussion of their statistical properties. Now, they are summarised interpreting identification algorithms used in the paper as estimators. These algorithms gave normalised, having the steady-state gain equal to 1.0000, mean values of models identified (frequency responses and the unit step responses) close to the corresponding normalised patterns. Additionally, the vari-

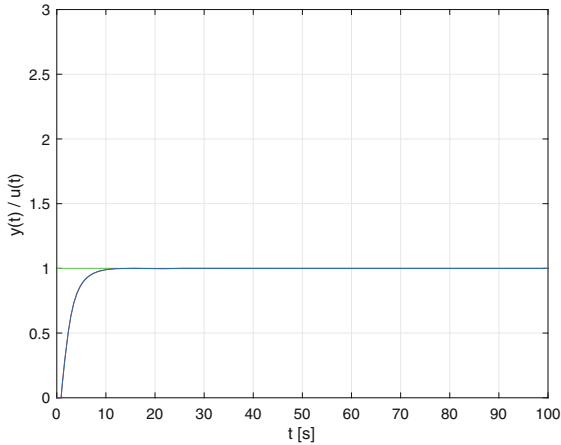


Fig. 10. Normalised mean value of 100 unit step responses calculated from the parametric models identified (blue line), the normalised pattern (red line) and the unit step excitation (green line) — $N = 262144$.

ance of the identification results obtained declines with the increasing number of the binary-valued observations processed.

Acknowledgments. The partial financial support of this research by The Polish Ministry of Science and Higher Education is gratefully acknowledged.

References

1. Figwer, J.: Synthesis and Simulation of Random Processes. Zeszyty Naukowe Politechniki Śląskiej, Seria Automatyka, Zeszyt nr 126: Gliwice, Poland (1999)
2. Figwer, J.: Continuous-time dynamic system identification with multisine random excitation revisited. *Archives Control Sci.* **20**, 123–139 (2010)
3. Figwer, J.: Process Identification by Means of Random Search Methods Illustrated with Simulation Examples (In Polish). Wydawnictwo Politechniki Śląskiej, Gliwice, Poland (2022)
4. Leong, A.S., Weyer, E., Nair, G.N.: Identification of FIR Systems with Binary Input and Output Observations. <https://doi.org/10.48550/arXiv.1809.06908>
5. Mathews, J.H., Fink, K.K.: Numerical Methods Using Matlab. Prentice Hall, Englewood Cliffs, New Jersey (2004)
6. Pouliquen, M., Pigeon, E., Gehan, O., Goudjil, A., Auber, R.: Impulse response identification from input/output binary measurements. *Automatica* **123** (2021). <https://doi.org/10.1016/j.automatica.2020.109307>



Time Series Identification Using Monte Carlo Method

Teresa Główka^(✉)

Department of Measurements and Control, Silesian University of Technology,
Akademicka 16, 44-100 Gliwice, Poland
TeresaGlowka@polsl.pl

Abstract. This paper focuses on time series (namely autoregressive, moving-average, and autoregressive-moving-average) models identification using one of the most universal Monte Carlo method, the Metropolis-Hasting algorithm. Theoretical formulation of a problem as well as some identification results are given.

Keywords: Monte Carlo methods · Metropolis-Hasting algorithm · time series identification · statistical signal processing

1 Introduction

The development of computational tools allowed for easy dispersal of computations, and therefore methods that can be easily dispersed in computer systems gain importance. Monte Carlo methods are an example of such methods.

Monte Carlo methods have been developed since 1940s and this process has accelerated since the 1990s due to the development of computer technology, see e.g. [1,2] and references therein. Since then, they have proven their usefulness in many fields of science and engineering as they can be applied in various integration, optimization and generally inference problems, for which a numerical solution is difficult or impossible to achieve (e.g. in high-dimensional problems).

The idea of Monte Carlo methods is very intuitive: instead of complicated numerical calculations, one should get multiple samples of the parameter he is looking for and average the result. This can be done by the physical experiment's repetitions or by a correct probabilistic description of the experiment, that enables the generation of a set of samples from the appropriate distribution using computers [1,2].

The basic Monte Carlo idea was soon developed by the rejection sampling algorithm and further by the Metropolis-Hastings algorithm [3,4]. This was the beginning of the Markov chain Monte Carlo approach. The next advance came with the Gibbs sampler algorithm in the 1980s. Since then, both the Metropolis-Hasting algorithm and the Gibbs sampler have been applied to several signal processing problems, like parameters estimation, blind identification, digital audio denoising, image reconstructions, and many others [1,2]. Many further versions of the Monte Carlo algorithms have been created, some of them applicable only in specific cases [1,2].

The discussion presented in this paper is complementary to the existing methods of time series identification, i.e. least squares, correlation, covariance, maximum likelihood, approximate maximum likelihood, minimum variance, see e.g. [5–8] spectral factorization, power spectral density approximation, whitening of equation error, minimisation of equation error energy methods [9]. The aim of this paper is to present the problem of time series models identification in a way that allows the use of the Metropolis-Hasting algorithm. Theoretical formulation of a problem is described in a way that allows the reader to apply it without turning to additional literature. This is followed by results of autoregressive, moving-average, and autoregressive-moving-average models identification, supplemented with some remarks on the algorithm parameterisation. To the best of our knowledge, there are no publications in the literature presenting this issue in a simple and comprehensive way. Moreover, it is not the author's intention to make a broad comparison to other existing methods due to the restrictive size requirements of this article. Such a comparison, as well as the use to identify other types of time series and objects, is under development and will be included in further works.

2 Monte Carlo Methods for Parameter Identification

The minimum mean squared error estimator $\hat{\boldsymbol{\theta}}$ of the parameters vector $\boldsymbol{\theta}$ can be expressed as [5]:

$$\hat{\boldsymbol{\theta}} = \int_{\boldsymbol{\Theta}} \boldsymbol{\theta} \pi(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}, \quad (1)$$

where $\pi(\boldsymbol{\theta}|\mathbf{y})$ is the conditional probability density function of $\boldsymbol{\theta}$, \mathbf{y} is the vector of observations, and $\boldsymbol{\Theta}$ is the multidimensional space of the parameters $\boldsymbol{\theta}$.

The most basic version of Monte Carlo method that allows for $\hat{\boldsymbol{\theta}}$ estimation can be formulated as follows: for $t = 1, \dots, T$ generate randomly $\boldsymbol{\theta}_t$, drawing them from the distribution given by the probability density function $\pi(\boldsymbol{\theta}|\mathbf{y})$, and next approximate Eq. (1) using the following mean of random samples $\boldsymbol{\theta}_t$ [1, 2]:

$$\hat{\boldsymbol{\theta}} = \frac{1}{T} \sum_{t=1}^T \boldsymbol{\theta}_t. \quad (2)$$

Here, parameterisation consists in selecting the number of drawn samples T .

In practice, generating samples directly from $\pi(\boldsymbol{\theta}|\mathbf{y})$ is often impossible. In this case one may use a modification known as the rejection sampling [1, 2].

In the rejection sampling algorithm, a simpler proposal density $q(\boldsymbol{\theta}|\mathbf{y})$ is used, for which the following inequality holds: $Mq(\boldsymbol{\theta}|\mathbf{y}) \geq \pi(\boldsymbol{\theta}|\mathbf{y})$, where M is a constant (i.e. M and $q(\cdot)$ are chosen to form an envelope $Mq(\cdot)$ of $\pi(\cdot)$). In this approach, in each t -th iteration of the algorithm, a sample candidate $\boldsymbol{\theta}'$ is drawn from $q(\boldsymbol{\theta}|\mathbf{y})$ and a random probability value u for such candidate is drawn from the standard uniform distribution $\mathcal{U}([0, 1])$. The sample $\boldsymbol{\theta}'$ is accepted (i.e. $\boldsymbol{\theta}_t = \boldsymbol{\theta}'$) and the algorithm goes to the next iteration, only if the acceptance probability $\alpha = \frac{\pi(\boldsymbol{\theta}'|\mathbf{y})}{Mq(\boldsymbol{\theta}'|\mathbf{y})} \geq u$. Otherwise this sample candidate is rejected and

the process of generating the new candidate $\boldsymbol{\theta}'$ for t -th iteration is repeated. Having all accepted samples $\boldsymbol{\theta}_t$, $t = 1, \dots, T$, the Eq. (2) is used to approximate the value of $\hat{\boldsymbol{\theta}}$. Parameterisation of this algorithm consists in selecting the number of drawn samples T , the proposal density $q(\boldsymbol{\theta}|\mathbf{y})$, and the constant M .

This procedure is most efficient if $Mq(\boldsymbol{\theta}|\mathbf{y})$ and $\pi(\boldsymbol{\theta}|\mathbf{y})$ are very similar in shape, because most of the generated candidates are accepted in this case. Otherwise, if they differ significantly, there is a need to generate much more candidates to obtain the set of T accepted samples.

2.1 Metropolis-Hasting Algorithm

The idea of Markov chain Monte Carlo methods is to generate the samples $\boldsymbol{\theta}_t$ from the desired distribution $\pi(\boldsymbol{\theta}|\mathbf{y})$ using ergodic Markov chain. In the Markov chain, a new sample $\boldsymbol{\theta}_t$ depends only on the value of the previous one $\boldsymbol{\theta}_{t-1}$ and not on the older ones (i.e. $\boldsymbol{\theta}_{t-2}, \dots, \boldsymbol{\theta}_0$). Note that this means that the consecutive values of samples are not independent.

The main algorithm within Markov chain Monte Carlo methods is the Metropolis-Hasting algorithm [1–4]. It can be interpreted as a generalisation of the rejection sampling algorithm, briefly described above. In the Metropolis-Hasting algorithm a proposal density $q(\cdot)$ is additionally dependent on the sample from previous iteration $\boldsymbol{\theta}_{t-1}$, and consequently the acceptance probability α for a new sample candidate $\boldsymbol{\theta}'$ is dependent on $\boldsymbol{\theta}_{t-1}$. In contrary to the rejection sampling, the draw of the sample candidate $\boldsymbol{\theta}'$ is not repeated if it is rejected. Consequently the new sample $\boldsymbol{\theta}_t$ is determined for each drawn candidate $\boldsymbol{\theta}'$, but in a different way. The value of the new sample $\boldsymbol{\theta}_t$ results from the candidate's acceptance: if the candidate $\boldsymbol{\theta}'$ is accepted, then the new sample becomes equal to it $\boldsymbol{\theta}_t = \boldsymbol{\theta}'$; if the candidate $\boldsymbol{\theta}'$ is rejected, then the new sample becomes equal to the previous one $\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1}$.

More specifically, the Metropolis-Hasting algorithm can be described as follows. In each t -th iteration, a sample candidate $\boldsymbol{\theta}'$ is drawn from $q(\boldsymbol{\theta}|\boldsymbol{\theta}_{t-1}, \mathbf{y})$ and a random probability value u for such candidate is drawn from the standard uniform distribution $\mathcal{U}([0, 1])$. The acceptance probability is calculated as:

$$\alpha = \min \left[1, \frac{\pi(\boldsymbol{\theta}'|\mathbf{y})q(\boldsymbol{\theta}_{t-1}|\boldsymbol{\theta}', \mathbf{y})}{\pi(\boldsymbol{\theta}_{t-1}|\mathbf{y})q(\boldsymbol{\theta}'|\boldsymbol{\theta}_{t-1}, \mathbf{y})} \right]. \quad (3)$$

If $\alpha \geq u$ then the candidate is accepted, i.e. $\boldsymbol{\theta}_t = \boldsymbol{\theta}'$, otherwise it is rejected and the previous value of the sample is kept, i.e. $\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1}$. Parameterisation of this algorithm consists in selecting the number of drawn samples T , the proposal $q(\boldsymbol{\theta}|\boldsymbol{\theta}_{t-1}, \mathbf{y})$ and the initial values of the chain $\boldsymbol{\theta}_0$. Additionally the number of initial samples T_b that are not taken to calculate the mean (so called the burn-in period) is usually chosen, so the Eq. (2) becomes

$$\hat{\boldsymbol{\theta}} = \frac{1}{T - T_b} \sum_{t=T_b+1}^T \boldsymbol{\theta}_t. \quad (4)$$

Note the important remarks: (1) the probability density function $\pi(\cdot)$ should be known up to a constant (so a probability $p(\cdot)$ can be used instead of $\pi(\cdot)$ if $\pi(\cdot) \propto p(\cdot)$), which is often the case [2]; (2) if the proposal density is symmetric, i.e. $q(\boldsymbol{\theta}_{t-1}|\boldsymbol{\theta}', \mathbf{y}) = q(\boldsymbol{\theta}'|\boldsymbol{\theta}_{t-1}, \mathbf{y})$, then the Eq. (3) simplifies to:

$$\alpha = \min \left[1, \frac{\pi(\boldsymbol{\theta}'|\mathbf{y})}{\pi(\boldsymbol{\theta}_{t-1}|\mathbf{y})} \right]. \quad (5)$$

There are two commonly used versions of the Metropolis-Hasting algorithm that can be found in the literature [2–4]:

- The random walk Metropolis-Hasting algorithm, in which the Markov chain is generated as a random walk process. A sample candidate $\boldsymbol{\theta}'$ is generated from the previous sample $\boldsymbol{\theta}_{t-1}$ modified by some additive random vector $\boldsymbol{\vartheta}'$, i.e. $\boldsymbol{\theta}' = \boldsymbol{\theta}_{t-1} + \boldsymbol{\vartheta}'$, where $\boldsymbol{\vartheta}'$ is drawn from a chosen multivariate probability density with mean values equal to zero. A common choice is the multivariate normal or uniform probability density. For example if $\boldsymbol{\vartheta}'$ is taken from $\mathcal{N}(\mathbf{0}, \mathbf{C}_{\boldsymbol{\vartheta}})$ (where $\mathbf{0}$ is the zero vector of mean values, $\mathbf{C}_{\boldsymbol{\vartheta}}$ is the covariance matrix), then the proposal density $q(\boldsymbol{\theta}'|\boldsymbol{\theta}_{t-1}, \mathbf{y}) = \mathcal{N}(\boldsymbol{\theta}_{t-1}, \mathbf{C}_{\boldsymbol{\vartheta}})$.
- The independent Metropolis-Hasting algorithm, where a sample candidate $\boldsymbol{\theta}'$ is generated independently from the previous sample $\boldsymbol{\theta}_{t-1}$, i.e. the proposal density $q(\boldsymbol{\theta}'|\boldsymbol{\theta}_{t-1}, \mathbf{y}) = q(\boldsymbol{\theta}'|\mathbf{y})$.

These algorithms behave differently: the random walk Metropolis-Hasting algorithm promotes the local exploration of a parameters space whereas the independent version enables more global search [2].

2.2 Time Series Identification with Metropolis-Hasting Algorithm

Autoregressive (AR), moving-average (MA), and autoregressive-moving-average (ARMA) models of the time series are discussed in the sequel. The general equation describing these models is [5]:

$$y(i) = H(z^{-1})e(i), \quad (6)$$

where i is the discrete time, $y(i)$ is the output of the time series (observations), $e(i)$ is the white noise excitation (unknown), z^{-1} is the backward time shift operator, i.e. $z^{-1}y(i) = y(i-1)$, and $H(z^{-1})$ is equal to $\frac{1}{A(z^{-1})}$ for AR model, $C(z^{-1})$ for MA model, and $\frac{C(z^{-1})}{A(z^{-1})}$ for ARMA model. Polynomials $A(z^{-1}) = 1 + a_1z^{-1} + \dots + a_{dA}z^{-dA}$ and $C(z^{-1}) = 1 + c_1z^{-1} + \dots + c_{dC}z^{-dC}$, both are assumed to be monic. Further it will be assumed that the white noise $e(i)$ is normally distributed, $e(i) \sim \mathcal{N}(0, \sigma^2)$, with the known variance σ^2 , and that the structure of the identified model is known (i.e. the orders dA and dC of the polynomials are known). Hence, the goal is to identify the parameters $\boldsymbol{\theta} = [a_1, \dots, a_{dA}, c_1, \dots, c_{dC}]^T$ given the vector \mathbf{y} of N observations $y(i)$, $i = 0, \dots, N-1$.

In the sequel, the formulation of the parameters θ identification task, that allows for using the Metropolis-Hasting algorithm, is specified. From Eq. (6) it is straightforward that:

$$e(i) = \frac{1}{H(z^{-1})}y(i). \quad (7)$$

As $e(i)$, $i = 0, \dots, N - 1$, are independent and normally distributed, the probability density function $\pi(\theta|\mathbf{y})$ is proportional to:

$$\pi(\theta|\mathbf{y}) \propto p(\theta|\mathbf{y}) = \prod_{i=0}^{N-1} \exp\left(-\frac{e(i)^2}{2\sigma^2}\right) = \exp\left(-\frac{1}{2\sigma^2} \sum_{i=0}^{N-1} e(i)^2\right). \quad (8)$$

For the sample θ_t the probability density function $p(\theta_t|\mathbf{y})$ is calculated from Eq. (8) substituting estimated excitation values $e_t(i)$ for $e(i)$. The values $e_t(i)$, $i = 0, \dots, N - 1$, are calculated from Eq. (7) substituting $H_t(z^{-1})$ for $H(z^{-1})$, where $H_t(z^{-1})$ is the estimate of $H(z^{-1})$ obtained for θ_t .

Now, the requirement for $H(z^{-1})$ to be stable and minimal phase (i.e. $\frac{1}{H(z^{-1})}$ is stable as well) is incorporated. It means, that the space Θ of the parameters $\theta = [a_1, \dots, a_{dA}, c_1, \dots, c_{dC}]^T$ is limited to those values of θ for which the roots of the polynomials $A(z^{-1})$ and $C(z^{-1})$ are inside the unit circle [5, 8]. So the indicator function $R(\theta)$ is defined, such that $R(\theta') = 1$ for all candidates θ' , for which the condition of $A(z^{-1})$ and $C(z^{-1})$ to have roots inside the unit circle is met, and $R(\theta') = 0$ otherwise. Finally, Eq. (8) for the probability density function of the new candidate θ' is modified:

$$\pi(\theta'|\mathbf{y}) \propto p(\theta'|\mathbf{y}) \times R(\theta') = R(\theta') \exp\left(-\frac{1}{2\sigma^2} \sum_{i=0}^{N-1} e'(i)^2\right), \quad (9)$$

where $e'(i)$ are estimated from Eq. (7) with $H'(z^{-1})$ substituted for $H(z^{-1})$, $H'(z^{-1})$ means the estimate of $H(z^{-1})$ obtained for θ' .

It should be emphasized that the same approach can be used to identify non-Gaussian time series. It can be done by modification of Eqs. (8) and (9) according to the non-Gaussian distribution function of $e(i)$.

3 Simulation Examples

This section contains some simulation results of parameters identification for AR, MA, and ARMA models using mostly the random walk Metropolis-Hasting algorithm, as well as some remarks on parameterisation of algorithm.

During the experiments the sequence $e(i)$, $i = 0, \dots, N - 1$, were drawn from $\mathcal{N}(0, \sigma^2)$, where $\sigma = 0.1$. The length of the data $N \in [50, 1000]$. Time series were simulated using the following exemplary equations, taken from [9]:

– in AR model identification:

$$y(i) = \frac{1}{1 - 1.5z^{-1} + 0.7z^{-2}}e(i), \quad (10)$$

– in MA model identification:

$$y(i) = (1 + 1.5z^{-1} + 0.7z^{-2})e(i), \quad (11)$$

– in ARMA model identification:

$$y(i) = \frac{1 - 0.8z^{-1}}{1 - 1.5z^{-1} + 0.7z^{-2}}e(i). \quad (12)$$

For random walk Metropolis-Hasting algorithm, the sample candidate θ' was drawn from $\mathcal{N}(\theta_{t-1}, \sigma_{\vartheta}\mathbf{I})$, and different σ_{ϑ} values were tested. Similarly, different initial values θ_0 of the random walk sequence were tested. The acceptance rate α in each iteration of the algorithm was calculated using Eq. (5), and Eqs. (8)–(9). The number of iterations of the algorithm was $T = 1000$, and the burn-in period was $T_b = 500$. So, the number of θ_t used to approximate $\hat{\theta}$, according to Eq. (4), was $T - T_b = 500$. The experiment was repeated 1000 times, each time for the same data and with the same parameterisation, in order to compute the mean and standard deviation of $\hat{\theta}$.

The identification results for the AR model are presented mainly to show the method's properties and the influence of parameterization. Similar results can be obtained in this case, for example, by the classical least squares approach. However, the advantage of the presented approach is the possibility of its use in the case of MA and ARMA series identification, which is not possible in a simple way for the classical least squares algorithm.

3.1 AR Model Identification

Figure 1 presents values of $\theta_t = [a_{1,t}, a_{2,t}]$ for $t = 1, \dots, T$, obtained during a single run of the random walk Metropolis-Hasting algorithm, for two different values of σ_{ϑ} . The rest of the parameters were: the length of the data $N = 100$, the initial values $\theta_0 = [0.5, -0.25]$ (so they were far from real values). For $\sigma_{\vartheta} = 0.1$

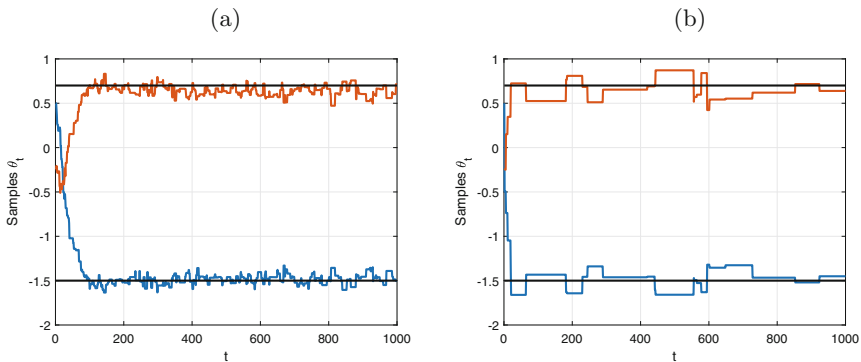


Fig. 1. The values of parameters $a_{1,t}$ (blue) and $a_{2,t}$ (red) during the single run of the algorithm obtained for (a) $\sigma_{\vartheta} = 0.1$, (b) $\sigma_{\vartheta} = 0.5$.

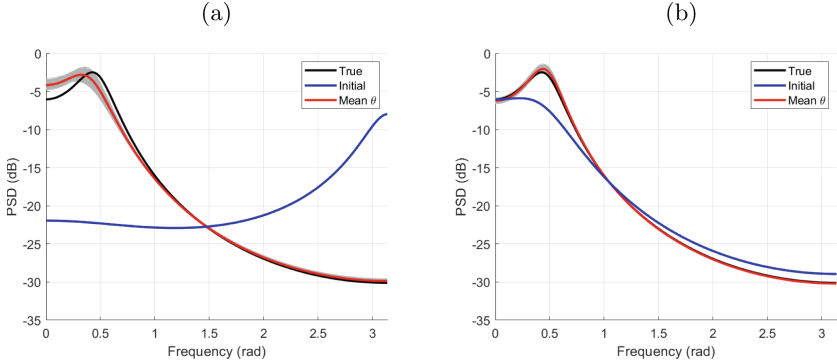


Fig. 2. PSDs obtained during 1000 repetitions of AR model identification experiment (gray lines), PSD obtained for mean of $\hat{\theta} = [\hat{a}_1, \hat{a}_2]$ from such experiments (red) and PSD obtained for initial values of parameters θ_0 (blue): (a) $N = 100$, $\theta_0 = [0.5, -0.25]$, $\sigma_\vartheta = 0.1$, (b) $N = 1000$, $\theta_0 = [-1.3, 0.5]$, $\sigma_\vartheta = 0.1$

(Fig. 1(a)) the number of accepted candidates θ' was c.a. 25%, for $\sigma_\vartheta = 0.5$ (Fig. 1(b)) it was less than 3%. However, we may observe, that moving towards the true values at the beginning of the waveforms was much faster for larger σ_ϑ . For small value $\sigma_\vartheta = 0.05$ (not depicted in Fig. 1), c.a. 50% candidates were accepted, for $\sigma_\vartheta = 0.01$ (also not shown) it was over 75%, but the follow-up properties were very poor in that case (the algorithm was not convergent in the assumed T_b time), and there was a high possibility of getting stuck in a local minimum. The rule of thumb, given in the literature, is to choose σ_ϑ to get somewhere between 25% and 40% accepted candidates [2].

The experiment as above was repeated 1000 times and the results are presented in frequency-domain in Fig. 2, as power spectral densities (PSDs), calculated from [5, 8]:

$$PSD(\omega) = 10 \log_{10} \left(\sigma^2 |\hat{H}(e^{-j\omega})|^2 \right), \quad (13)$$

where $\hat{H}(e^{-j\omega})$ was the estimate $\hat{H}(z^{-1})$ obtained for parameters $\hat{\theta}$ with substitution $z^{-1} = e^{-j\omega}$ and $\omega \in [0, \pi)$ was the relative frequency. All PSDs calculated for all $\hat{\theta}$ (recall that for a single run, $\hat{\theta}$ was calculated as the mean of θ_{501} to θ_{1000}) in 1000 runs of the random walk Metropolis-Hasting algorithm are depicted as gray lines. PSD calculated for the mean of all 1000 values of $\hat{\theta}$ is shown as red line, and PSD of the initial values θ_0 (the same in all 1000 runs) as blue line. Figure 2(a) presents the results for $N = 100$ (short data sequence), $\theta_0 = [0.5, -0.25]$ (far from true values), Fig. 2(b) shows results for $N = 1000$ (long data sequence), $\theta_0 = [-1.3, 0.5]$ (close to true values). For both experiments $\sigma_\vartheta = 0.1$.

Repeating the experiment allows to calculate the mean and standard deviation of $\hat{\theta}$, as shown in Table 1 for various N and θ_0 . From Fig. 2 and Table 1 one may conclude that increasing N the results are closer to true values and their standard deviation decreases. Comparing what happened if the initial values

were close to true ones (column two and three in Table 1) or far from them (column four and five in Table 1), one may see that the results for $N = 50$ and 100 are almost identical. However, for $N > 100$ and the initial values far from true ones the algorithm often could not move from the initial values giving totally erroneous results, so these results are not written in Table 1. This reflects the intuition that for random walk algorithms, the right choice of initial values is very important.

It is not the case in the independent Metropolis-Hasting algorithm, where the new candidate is not drawn from the surrounding of the previous sample, but in the whole sample space according to the chosen proposal density. 1000 runs of independent Metropolis-Hasting algorithm with the uniform proposal $\mathcal{U}[-2, 2]$ for a_1 and $\mathcal{U}[-1, 1]$ for a_2 , and for $N = 1000$ gave the results -1.5166 (0.0514) for a_1 and 0.7207 (0.0509) for a_2 . A much larger standard deviation of estimates can be observed comparing to the random walk with “good” initial values. However, independent algorithm in each run gave some reasonable result in surrounding of true values, in contrary to the random walk with “bad” initial values. Therefore the conclusion can be drawn to start with a single run of the independent Metropolis-Hasting algorithm, and treat its results as the initial values of the random walk Metropolis-Hasting algorithm.

3.2 MA and ARMA Model Identification

Figures 3 and 4 shows the results of 1000 repetitions of identification for MA and ARMA models, respectively. The conclusions that can be drawn from these results are generally similar to those presented for AR model identification. However, it can be seen that the results for ARMA model are less accurate than those for AR and MA models. Especially it is seen for $N = 100$, see Fig. 4(a), where it can be observed that a number of single runs of the algorithm gave unsatisfactory results, because the algorithm did not converge in a single run time (although the average is quite good, meaning only a few runs were non-convergent). To

Table 1. Statistical results of AR model identification calculated over 1000 runs of algorithm: mean of parameters estimates and their standard deviation in parentheses, obtained for five different N and two sets of initial values.

N	a_1	a_2	a_1	a_2
50	-1.3815 (0.0296)	0.5009 (0.0302)	-1.3831 (0.0296)	0.5027 (0.0306)
100	-1.4772 (0.0164)	0.6385 (0.0165)	-1.4758 (0.0158)	0.6374 (0.0162)
200	-1.4931 (0.0098)	0.7123 (0.0101)		
500	-1.5051 (0.0069)	0.7245 (0.0070)		
1000	-1.5152 (0.0066)	0.7195 (0.0067)		
True	-1.5	0.7	-1.5	0.7
Initial	-1.3	0.5	0.5	-0.25

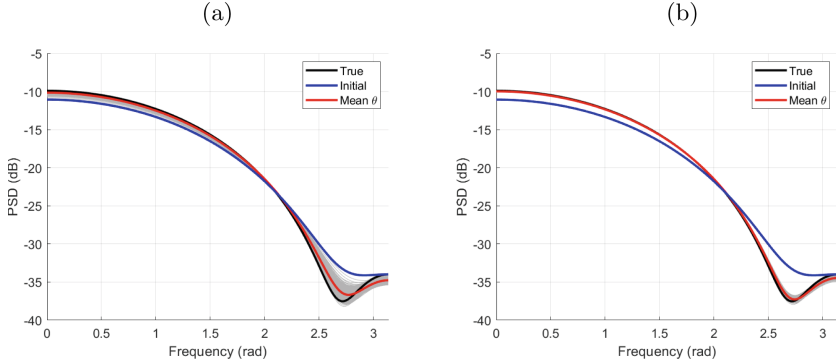


Fig. 3. PSDs obtained during 1000 repetitions of MA model identification experiment (gray lines), PSD obtained for mean of $\hat{\theta}$ from such experiments (red) and PSD obtained for initial values of parameters $\theta_0 = [1.3, 0.5]$ (blue): (a) $N = 100$, $\sigma_{\vartheta} = 0.1$, (b) $N = 1000$, $\sigma_{\vartheta} = 0.1$

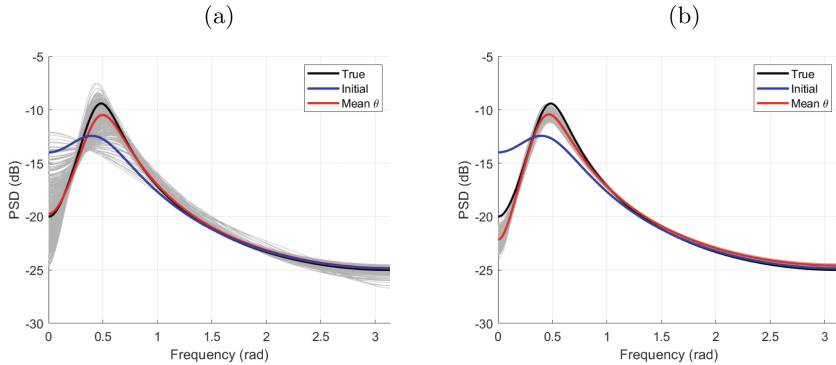


Fig. 4. PSDs obtained during 1000 repetitions of ARMA model identification experiment (gray lines), PSD obtained for mean of $\hat{\theta}$ from such experiments (red) and PSD obtained for initial values of parameters $\theta_0 = [-1.3, 0.5, 0.6]$ (blue): (a) $N = 100$, $\sigma_{\vartheta} = 0.1$, (b) $N = 1000$, $\sigma_{\vartheta} = 0.1$

overcome this problem, the number of the iteration T (and proportionally T_b) in a single run should be increased.

4 Conclusions and Final Remarks

The problem of time series models identification using one of the Markov chain Monte Carlo method, the Metropolis-Hasting algorithm, has been raised in the paper. The problem of autoregressive, moving-average, and autoregressive-moving-average models identification has been formulated and the given examples of identification results for different lengths of data have shown that increasing the data length allows not only to improve the identification results but also

to reduce their standard deviations calculated over multiple runs of the algorithm. The influence of the initial values on the identification results using random walk Metropolis-Hasting algorithm was shown. The proposal to determine the initial values for random walk Metropolis-Hasting algorithm from a single run of independent Metropolis-Hasting algorithm has been suggested.

Finally some remarks can be drawn. First, for simplicity there has been assumed that the variance of the white noise excitation was known. If it is not the case, then one of the polynomials $A(z^{-1})$ or $C(z^{-1})$ should not be assumed to be monic, and the additional identified parameter a_0 or c_0 , after being pulled out in front of the parentheses, can be used to determine the unknown variance of the white noise excitation. Second, in general the parameters space can have a lot of local minima for larger degrees of the polynomials $A(z^{-1})$ and $C(z^{-1})$. This can strongly affect the convergence of the algorithm. Moreover it can lead to the situation, where the parameters corresponding to different local minima are averaged, giving totally erroneous results. However, there are some approaches proposed in the literature that deal with this problem [9]. This will also be the subject of further research.

Acknowledgement. The research reported in this paper has been supported by State Budget for Science, Poland: Silesian University of Technology grant no. 02/050/BK_23/0032.

References

1. Robert, C.P., Casella, G.: Monte Carlo Statistical Methods. Springer, New York (2004)
2. Luengo, D., Martino, L., Bugallo, M., Elvira, V., Särkkä, S.: A survey of Monte Carlo methods for parameter estimation. EURASIP J. Adv. Signal Process. **2020**(1), 1–62 (2020). <https://doi.org/10.1186/s13634-020-00675-6>
3. Hastings, W.K.: Monte Carlo sampling methods using Markov chains and their applications. Biometrika **57**(1), 97–109 (1970)
4. Hitchcock, D.B.: A history of the Metropolis-Hastings algorithm. Am. Stat. **57**(4), 254–257 (2003)
5. Kay, S.M.: Modern Spectral Estimation: Theory and Applications. Prentice-Hall, New Jersey (1988)
6. Lütkepohl, H.: Introduction to Multiple Time Series Analysis. Springer-Verlag, New York (1993)
7. Niederliński A., Kasprzyk J., Figwer J.: MULTI-EDIP Analizator wielowymiarowych sygnałów i obiektów. Wydawnictwo Politechniki Śląskiej, Skrypt nr 2017, Gliwice (1997)
8. Marple, S.L.: Digital Spectral Analysis with Applications. Prentice Hall, Englewood Cliffs (1987)
9. Figwer, J.: Process Identification by Means of Random Search Methods Illustrated with Simulation Examples (In Polish). Wydawnictwo Politechniki Śląskiej, Gliwice, Poland (2022)



Calculation Method of the Centrifugal Pump Flow Rate Based on Its Nominal Data and Pump Head Increase

Uliana Nykolyn^(✉) and Petro Nykolyn

Ivano-Frankivsk National Technical University of Oil and Gas, Ivano-Frankivsk, Ukraine
uliana.nykolyn@nung.edu.ua

Abstract. It has been developed the method to calculate the volumetric flow rate for an energy audit investigation of centrifugal pumps with electric drives based on a previously created mathematical model using the principles of electrohydraulic analogy. The energy efficiency of the centrifugal pump was assessed through the implementation of an informative parameter - the load angle. This angle characterizes the operation mode of the power unit. The flow rate of the centrifugal pump due to the head gradient is calculated. The mutual dependence of the pump load angle on its volumetric flow rate is shown. This approach makes it possible to reduce financial costs and technically simplifies the energy audit, as there is no need to purchase an ultrasonic flow meter and its installation. The head characteristics of one of the centrifugal pumps were constructed using the proposed method. In addition, the developed method makes it possible to evaluate the energy efficiency of centrifugal pumps of small and medium power.

Keyword: Centrifugal pump · Energy efficiency · Load angle · Efficiency · Power consumption

1 Introduction

Centrifugal pumps make up a significant part of the industrial load and energy infrastructure of cities, which consume up to 10% of generated electrical energy [1]. This makes it necessary to take measures to increase the efficiency of their work. The problem also lies in the fact that most low-power pumps work without an audit of their efficiency, since only large-volume units are equipped with flow meters [2]. The operation of units with a low efficiency factor is due to inconsistency between the technical characteristics of the electric drive and hydraulic parts of the unit, untimely repairs, long-term work in a mode that is far from optimal.

Usually, in order to evaluate the efficiency of the pumps, it is necessary to have such data as the flow rate, the pump head at the inlet and outlet, the coefficient of speed, and the design parameters of the unit. The greatest difficulty is to obtain flow data due to the lack of a flow meter. Different types of flowmeters are used to determine the flow rate, including turbine flowmeters, electromagnetic and ultrasonic flowmeters. The main disadvantages of these devices are their installation in the section of the pipeline, their

constant maintenance, increased requirements for the characteristics of the liquid, the price, and the presence of errors in measurements is also significant.

Determining the necessary parameters by calculation, in particular the flow rate, based on mathematical modeling of the processes that occur in the centrifugal unit, will make it possible to avoid a significant part of the above mentioned disadvantages when evaluation of the efficiency of its work is done.

2 Methods Overview

A literature review shows that it has been developed the numerous evaluating methods of the centrifugal pump efficiency. For example, the author's team of Finnish scientists under the leadership of Tero Ahonen [3] has proposed a new method of estimating energy consumption which uses the current values of only three parameters of the pump mode: differential head H_D , the drive motor consumed power N_C , and the rotation speed of the impeller n . However, they do not take into account the manufacturer's characteristics of the pump. Besides the proposed method is not accurate enough, because in different operating modes of the aggregates it gives an error of 7% to 20%.

Recently, the Yatesmeter method [4] based on the thermodynamic approach has gained popularity. The essence of this method is to determine the efficiency of the centrifugal pump based on the temperature difference of the working fluid at its inlet and outlet. But the disadvantage of this method is the impossibility of taking into account all the heat released during the work, as well as the difficulty of using thermal sensors due to the correct location of their installation, sensitivity to external influences of an electromagnetic and thermal nature, special requirements for the pipeline surface. Highly sensitive sensors with high resolution are too expensive and other sensors are ineffective due to large error [5–7].

To solve this problem, mathematical models of the centrifugal pump are increasingly used today, which make it possible to determine the consumption load of the pump indirectly without using of expensive measuring devices. The prerequisite for the theoretical construction of adequate energy characteristics of the centrifugal pump was the creation of the mathematical model of the centrifugal pump based on the electrohydraulic analogy method [8] which makes it possible to construct the operating characteristics of the pump according to its design data taking into account such physical properties of the working fluid as its density and viscosity. The detailed substitution scheme of a centrifugal pump operating with concentrated complex parameters is given in [8].

3 Main Part

The application of the electrohydraulic analogy method is based on the implementation of the electric circuit theory for the mathematical description of energy conversion processes in centrifugal pumps. The electrical equations are transformed into the corresponding hydromechanical dependences, which makes it possible to analyze the efficiency of the centrifugal pump.

To compare the characteristics of different types vane pumps in hydromechanics the speed coefficient n_S is usually used, which is determined by the expression (1)

$$n_S = 3,65n^{nom} \sqrt[4]{\left(\frac{Q_D^{nom}}{M}\right)^2 \left(\frac{L}{H_D^{nom}}\right)^2}, \quad (1)$$

where n^{nom} - nominal speed of rotation; Q_D^{nom} - nominal flow rate; H_D^{nom} - nominal pump head; L , M - respectively, the number of working pressure stages and working flows of the pump.

In electromechanics the electromagnetic load angle θ is used to analyze the reliability and efficiency of synchronous electric machines operating modes [9]. This is the angle between the vector of the electromotive force of the machine (electric generator or engine) and the vector of voltage on its clamps, which changes proportionally depending on the moment on the shaft. Usually its value is 60–90 degrees.

The electrohydraulic analogy method and its model were developed by professor V. Kostyshyn (Ukraine) [8]. This model is based on the creation of an electrical substitution scheme of the pump. The model consists of the active resistances to reflect losses and inertial parameters to characterize useful energy conversion processes. As it was mentioned above, this model allows to build the pump operating characteristics using design data and fluid density and viscosity. The proposed technique allows to analyze the operation of a classic design centrifugal pump which is driven by an electric motor.

Based on the centrifugal pump mathematical model [8], the calculation of the load angle γ_{calc}^{nom} using passport data is proposed

$$\gamma_{calc}^{nom} = \pi \left(1 - \frac{k_{Dcalc}}{H_{*0} \mu_H \eta_Z^{nom}}\right) \mu_Q \eta_{vol}^{nom}, \quad (2)$$

where

k_{Dcalc} - the ratio of the outer and inner diameter of the impeller; H_{*0} - the relative value of the pump head in the no-load mode; μ_H , μ_Q - respectively, coefficients of head reduction and volume compression of the working flow due to the finite number of blades; η_Z^{nom} , η_{vol}^{nom} - respectively hydraulic and volumetric efficiency of the pump.

It should be mentioned that the value η_Z^{nom} has been calculated using electrohydraulic analogy method [page 148, 9]. This parameter is little variable; it changes in the range from 1 to 0.95 on the load change interval from 0 to the nominal value.

The final simplification of the substitute circuit of the centrifugal pump allows you to determine the calculated load angle through the equivalent inertial resistance of the pump circuit

$$\gamma_{calc}^{nom} \approx 2 \arctg(x_{*total}), \quad (3)$$

where x_{*total} - total reactance of the equivalent circuit of the pump.

Let's combine the specific speed coefficient n_S of the machine and its load angle in the expression (4)

$$\gamma_{calc}^{nom} = 0,475 \left(1 + \frac{n_S}{100} \right), \quad (4)$$

We will calculate the main energy parameters to determine the power consumption and energy efficiency of the centrifugal pump. For example, consider the main pump HM-7000–210, the parameters of which are given in the Table 1. The Table 1 contains the nominal passport values of the pump according to the manufacturer's instructions, as well as the calculated electro-hydraulic parameters [9].

Pump NM-7000–210 is a horizontal centrifugal single-stage pump for pumping oil, for main pipelines, numbers of working streams - 2, outer diameter 0.465 m, 8 impeller blades, blade thickness - 0.004 m, multiplicity of outer and inner diameters - 1.95, blade angle 210^0 , shaft sealing - end, specific speed coefficient $n_S = 195$, electric drive power 5000 kW.

Table 1. Nominal and calculated parameters of the pump HM-7000–210

H_D^{nom} m	Q_D^{nom} , m ³ /sec	n^{nom} , rpm	η_{nom}	γ_{calc}^{nom}	x_{*total}	n_S	μ_Q	μ_H	N_C^{nom} , kW
210	1,944	3000	0,87	1,38	0,892	195	0,897	0,831	4604

We will determine the flow of the centrifugal pump based on the experimentally obtained head on the outlet pipe and the calculated load angle using the expression (5)

$$Q_{*D} = \sqrt{\frac{\left(\frac{\gamma_{calc}^{nom}}{\sin(\gamma_{calc}^{nom})} - H_{*D} \right)}{\left(\frac{\gamma_{calc}^{nom}}{\sin(\gamma_{calc}^{nom})} - 1 \right)}}, \quad (5)$$

where H_{*D} , Q_{*D} - respectively, the value of the head and the flow rate at the pump outlet nozzle in the system of relative units, where the nominal pump parameters are taken as the basic ones [8].

It is advisable to make calculations in the system of relative units. The basic parameters are determined by (6) and (7)

$$\left. \begin{aligned} H_{base} &= H_D^{nom}, \\ Q_{base} &= Q_D^{nom}. \end{aligned} \right\} \quad (6)$$

And basic power N_{base} and basic impedance Z_{base}

$$\left. \begin{aligned} N_{base} &= \rho \cdot g \cdot H_{base} Q_{base}, \\ Z_{base} &= \frac{\rho \cdot g \cdot H_{base}}{Q_D^{nom}}. \end{aligned} \right\} \quad (7)$$

and the relative values of the parameters of the mode and the substitute scheme will be denoted by an index “*”

$$\left. \begin{aligned} H_* &= \frac{H}{H_{base}}; Q_* = \frac{Q}{Q_{base}}, \\ N_* &= \frac{N}{N_{base}}; Z_* = \frac{Z}{Z_{base}} \end{aligned} \right\}. \quad (8)$$

In particular, in the system of relative units (for an incompressible fluid), the dimensionless values of pressure P_* and head H_* are equal to each other

$$P_* = \frac{\rho \cdot g \cdot H}{\rho \cdot g \cdot H_{base}} = \frac{H}{H_{base}} = H_*. \quad (9)$$

Similarly, we determine the power consumed from the shaft of the drive motor of pump

$$N_{*C} = \frac{1}{\eta_{nom}} \left[1 + \left(3 - 2 \frac{\gamma_{calc}^{nom}}{\sin(\gamma_{calc}^{nom})} \right) \left(\sqrt{\frac{\left(\frac{\gamma_{calc}^{nom}}{\sin(\gamma_{calc}^{nom})} - H_{*D} \right)}{\left(\frac{\gamma_{calc}^{nom}}{\sin(\gamma_{calc}^{nom})} - 1 \right)}} - 1 \right) \right], \quad (10)$$

where η_{nom} – nominal efficiency of the pump.

The total efficiency of the centrifugal unit will be

$$\eta = \frac{\sin(\gamma_{calc}^{nom} Q_{*D})}{\sin(\gamma_{calc}^{nom}) + [(Q_{*D} - 1)\gamma_{calc}^{nom}] \cos(\gamma_{calc}^{nom})}. \quad (11)$$

Equations (4), (5) and (10) were obtained as a result of mathematical derivations and simplifications during the analysis of physical processes based on the mathematical model of real centrifugal pump [8].

As you can see, Eqs. (2), (5), (10), (11) are defined for the nominal point of the pump characteristic curve. With the change of working flows and heads both inside and the outlet of the pump, the electrical impedances (active resistances and reactances) of its complex substitution scheme, which was created on the basis of an electrohydraulic analogy, will also change. To construct other mode points, all parameters are changed accordingly.

4 Result Analysis

Usually, the operating characteristics of the centrifugal pump determine the dependence of the parameters H_D , N_C and η on the flow rate Q_D . In our case, the power characteristics of the main pump HM-7000–210 shown in Fig. 1 are plotted relative to the head value H_{*D} . This fact reflects the main idea of the proposed method – to express the change in the energy parameters of the pump as function of the head value because we are trying to do without a flowmeter, using only the manometers.

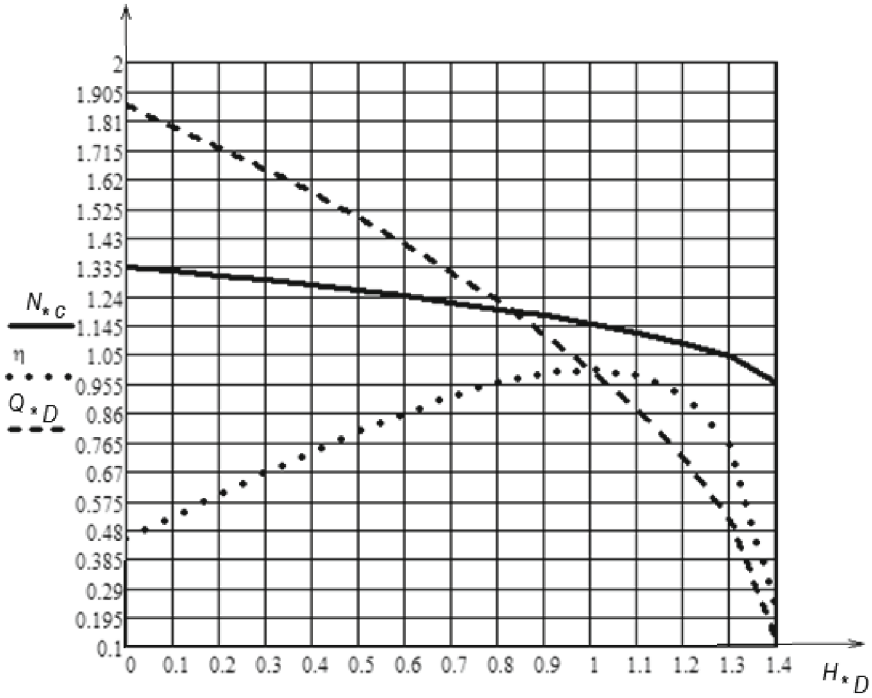


Fig. 1. Head and power characteristics of the pump HM-7000-210

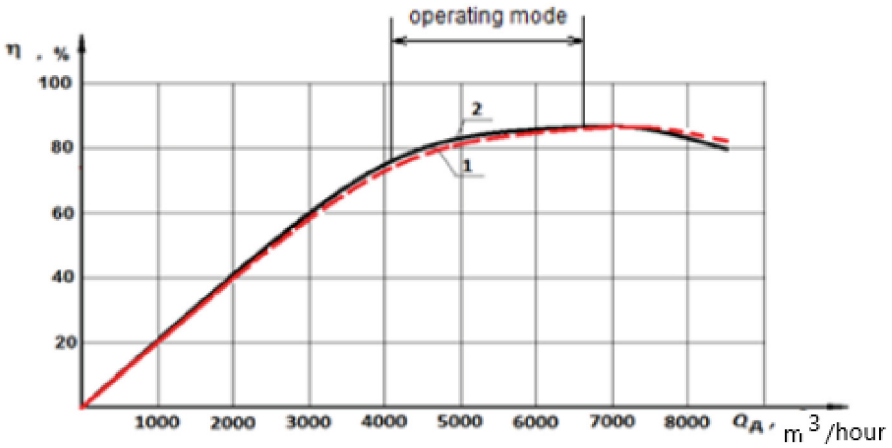


Fig. 2. The dependence of the efficiency characteristics on the flow rate of the HM 7000-210 pump, calculated using the presented model (line 1) and obtained experimentally (line 2)

Figure 2 illustrates a good match between the efficiency characteristics of the main centrifugal pump HM-7000-210 calculated using the mathematical model and obtained experimentally.

In fact, at the objects of the oil industry in Ukraine, in particular “Druzhba” oil main pipeline or the oil group collection in Dolyna (Ivano-Frankivsk region, Ukraine), the operation of the centrifugal pumps is carried out with a load factor of 0.6 - 0.9. The extreme limits of pump operation are unlikely in real operating conditions. Therefore, in the working interval of the pump (shown in Fig. 2), the experimentally recorded curve and the curve calculated by the model differ little (up to 10%), which indicates the adequacy of the proposed calculation method.

The degree of wear of various structural parts (for example, seals, blades of the impeller, roughness of the inner surface of the centrifugal pumps) is different and affects the amount of its head drop in different ways. In pumps with a higher degree of wear, the head gradient between its inlet and outlet will differ from a pump with a lower degree of wear. There cannot be universal method of analysis. But in the proposed method, the head at the pump outlet is one of the informative parameters for determining the flow rate, power and efficiency.

Thus, the evaluation of the efficiency of the centrifugal pump operation requires experimental determination only one technological parameter of the pump mode, namely the differential head H_D . With the help of standard manometers.

5 Conclusions

1. Based on the electro-hydraulic model of the centrifugal pump, the main energy parameters were calculated without taking into account the flow meter.
2. The adequacy of the proposed method is confirmed by the convergence of the energy dependences of the pump obtained by calculation and experiment.
3. The presented method makes it possible to assess the energy efficiency of centrifugal pump using less numbers of experimentally obtained data adequately.

References

1. Waide, P., Brunner, C.: Energy-efficiency policy opportunities for electric motor-driven systems, International Energy Agency (2011)
2. European Commission: Communication from the Commission to the European Parliament and the Council: Energy efficiency and its contribution to energy security and the 2030 framework for climate and energy policy. Brussels **23**, 7 (2014)
3. Specific speed-based pump flow rate estimator for large-scale and long-term energy efficiency auditing Santeri Poyhonen & Tero Ahonen & Jero Ahola & Pekka Punnonen & Simo Hammo & Lauri Nygren. Accessed 13 Nov 2018
4. <http://www.atap.ca/yatesmeter.html>. Accessed 16 Nov 2020
5. <http://ena.lp.edu.ua/bitstream/ntb/21644/1/12-64-67.pdf>. Accessed 20 March 2021
6. <http://learn.ztu.edu.ua/mod/resource/view.php?id=5712>. Accessed 30 Sept 2021
7. http://www.kdu.edu.ua/statti/Tezi/Tezi_2012/132.pdf. Accessed 06 Dec 2020
8. Костишин, В.С.: Моделювання режимів роботи відцентрових насосів на основі електрогідравлічної аналогії. Івано-Франківськ. Факел (2000)
9. Яцун, М.А.: Електричні машини: Навчальний посібник. Львів: Видавництво Національного університету "Львівська політехніка" (2001)



A Majorization-Minimization Algorithm for Optimal Sensor Location in Distributed Parameter Systems

Dariusz Uciński^(✉)

Institute of Control and Computation Engineering, University of Zielona Góra,
ul. prof. Z. Szafrana 2, 65-516 Zielona Góra, Poland
d.ucinski@issi.uz.zgora.pl

Abstract. The wide availability of effective and efficient convex optimization algorithms makes convex relaxation of optimum sensor location problems enjoy high popularity. The main drawback of this approach, however, is that the performance gap between the optimal solution of the original combinatorial problem and the heuristic solution of the corresponding relaxed continuous problem can hardly be neglected. As a countermeasure, the original design criterion is often extended by addition of some kind of sparsity-enforcing penalty term. Unfortunately, the appealing problem convexity is then lost and the question of how to control the influence of this penalty so as not to excessively deteriorate the optimal relaxed solution remains open. This is why an alternative problem formulation is proposed here, in which the sparsity-promoting term is directly minimized subject to the constraint that the efficiency of the sensor configuration with respect to the employed design criterion is no less than an arbitrarily set threshold. This brings the degree of optimality of the produced solution under direct control. An efficient majorization-minimization algorithm is then employed in combination with generalized simplicial decomposition to produce sparse relaxed solutions. The attendant computations boil down to solving a sequence of low-dimensional convex optimization problems, which constitutes a clear advantage of the proposed technique.

1 Introduction

Distributed parameter systems (DPSs) form a class of dynamic systems whose states depend on both time and space. Their traditional description are partial differential equations (PDEs). In most cases, not all physical parameters underlying such models can be directly measured and they have to be estimated via calibration yielding the best fit of the model output to the observations of the actual system which are provided by measurement sensors. But the number of sensors is most often limited, which causes the problem of where to locate them so as to collect the most valuable information about the parameters.

The classical approach to optimal sensor location consists in formulating it as an optimization problem employing various design criteria defined on the Fisher

information matrix (FIM) associated with the estimated parameters. Comprehensive overviews of the works published in this area are contained in the monographs [12, 15, 19].

It goes without saying that the interest in this problem has increased rapidly due to the growing popularity of sensor networks. More and more complex settings have been investigated to meet needs created by a variety of practical scenarios. They have mainly been focused on properly addressing the ill-posedness inherent in problems with large (or even infinite) dimensions of the parameter space [1, 2, 8, 9].

The number of sensors is usually imposed by the available experimental budget. Therefore, most techniques boil down to the selection of optimal sensor locations from a finite (but possibly very large) set of candidate locations. Note that the problem of assigning sensors to specific spatial locations can equivalently be interpreted in terms of activating an optimal subset of all the available sensors deployed in the spatial area (the non-activated sensors remain dormant). This framework is typical of the measurement regime characteristic for modern sensor networks.

A severe difficulty in selecting an optimal subset of gauged sites from among a given set of candidate sites is the combinatorial nature of this optimization problem. As the cardinalities of those sets increase, the exhaustive search of the search space quickly becomes computationally intractable. This stimulated attempts to solve this problem in a smarter manner. For problems with low or moderate dimensionalities, in [23] a branch-and-bound method was set forth, which most often drastically reduces the search space to produce an optimal integral solution. In turn, for large-scale observation networks, the existing approaches replace the original NP-hard combinatorial problem with its convex relaxation in the form of a convex programming problem. This paves the way for application of interior-point methods [6, 11] or polyhedral approximation methods [10, 21, 23].

A major drawback of convex relaxation is the necessity of transforming the optimal relaxed solution into an acceptable solution of the original combinatorial problem. This is by no mean trivial and, when done carelessly, may make the performance gap between both the solutions quite wide. In recent years, handling the problem with various sparsity-enforcing penalty terms, which is called sparsity control, have won popularity [1]. Although as a result of their addition to the original design criteria the problem convexity is lost, this property can be easily retrieved by resorting to iterations of the majorization-minimization scheme, cf. [18]. One major disadvantage of this powerful scheme, however, is that it is not clear how to control the impact of this penalty so as not to depart from the relaxed solution too much.

The main contribution of this work consists in establishing an alternative approach which explicitly controls the quality of the sparsified solution in terms of the original design criterion. It focuses on directly minimizing the sparsity-enforcing penalty within the set of relaxed solutions which are allowed to deteriorate the optimal relaxed solution in terms of the original design criterion by no more than an arbitrarily set amount. The proposed technique reduces to

a sequence of nonlinearly constrained convex optimization problems which are solved using an extremely fast generalized simplicial decomposition. As a result, a relatively simple and efficient technique of postprocessing relaxed solutions is proposed.

2 Optimal Sensor Location and Its Convex Relaxation

Consider a spatiotemporal system whose scalar state is given by the solution y to a deterministic partial differential equation (PDE) accompanied by the appropriate boundary and initial conditions. The PDE is defined on a bounded spatial domain $\Omega \subset \mathbb{R}^d$ ($d \leq 3$) with a boundary $\partial\Omega$ and a bounded time interval $T = (0, t_f]$. It is specified up to $\boldsymbol{\theta} \in \mathbb{R}^m$, a vector of unknown parameters which are to be estimated from noisy observations of the state. These observations are going to be made by n pointwise sensors at given time instants $t_1, \dots, t_K \in T$.

The sensor locations are to be selected from among $N > n$ candidate sites $\mathbf{x}^1, \dots, \mathbf{x}^N \in \bar{\Omega} := \Omega \cup \partial\Omega$. Their measurements are modelled as

$$z_{j,k} = y(\mathbf{x}^{i_j}, t_k; \bar{\boldsymbol{\theta}}) + \varepsilon_{i_j,k} \quad (1)$$

for $j = 1, \dots, n$ and $k = 1, \dots, K$, where $y(\mathbf{x}, t; \boldsymbol{\theta})$ stands for the state at a spatial point $\mathbf{x} \in \bar{\Omega}$ and a time instant $t \in \bar{T} := [0, t_f]$, evaluated for a given parameter $\boldsymbol{\theta}$. Here $\bar{\boldsymbol{\theta}}$ signifies the vector of ‘true’ values of the unknown parameters, $i_1, \dots, i_n \in \{1, \dots, N\}$ are the indices of the gauged sites, and the $\varepsilon_{i_j,k}$ ’s are independent normally-distributed random errors with zero mean and constant variance σ^2 .

When a maximum-likelihood or a maximum *a posteriori* estimate $\hat{\boldsymbol{\theta}}$ of $\bar{\boldsymbol{\theta}}$ is produced, its accuracy is characterized by the Fisher information matrix (FIM), cf. [3]

$$\mathbf{M}(\mathbf{v}) = \mathbf{M}_0 + \sum_{i=1}^N v_i \mathbf{M}_i, \quad (2)$$

in which

$$\mathbf{M}_i = \frac{1}{\sigma^2} \sum_{k=1}^K \mathbf{g}(\mathbf{x}^i, t_k) \mathbf{g}^\top(\mathbf{x}^i, t_k), \quad i = 1, \dots, n, \quad (3)$$

with $\mathbf{g}(\mathbf{x}, t) = \nabla_{\boldsymbol{\theta}} y(\mathbf{x}, t; \boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^0}$ standing for the first-order sensitivity vector of the state variable evaluated at a prior estimate $\boldsymbol{\theta}^0$ of the ‘true’ vector $\bar{\boldsymbol{\theta}}$. If \mathbf{M}_0 is present, its inverse is the known covariance matrix of the prior density of $\boldsymbol{\theta}$. Furthermore, $\mathbf{v} = (v_1, \dots, v_N)$ and v_i is the binary indicator variable equal to 1 or 0 depending on whether or not a sensor resides at site \mathbf{x}^i .

The amount of information about unknown parameters provided by a sensor location represented as \mathbf{v} can be quantified by a design criterion related to the confidence ellipsoid, i.e., a highest probability region for the unknown parameters. The most common option here is the D-optimality criterion

$$\Phi_D(\mathbf{M}) = \det^{1/m}(\mathbf{M}). \quad (4)$$

This is because its maximization is equivalent to minimizing the volume of the confidence ellipsoid. As a result, the spread of the estimates of the unknown parameters around their ‘true’ values is minimized.

Without loss of generality, in what follows we assume that the design criterion is just (4). Other criteria [14] can be treated in much the same way.

Consequently, we can formulate the following discrete-optimization problem:

Problem 1. Find $\mathbf{v}_{\text{bin}}^* \in \mathcal{V}_{\text{bin}} := \{\mathbf{v} \in \{0, 1\}^N : \mathbf{1}^\top \mathbf{v} = n\}$ to maximize $\Phi_{\text{D}}(\mathbf{M}(\mathbf{v}))$. Each minimizer $\mathbf{v}_{\text{bin}}^*$ is called a D-optimum *exact* design.

It is easy to see that the exhaustive search for $\mathbf{v}_{\text{bin}}^*$ through evaluating $\Phi_{\text{D}}(\mathbf{M}(\mathbf{v}))$ for each possible choice of gauged sites is computationally intractable even for moderate N . A common remedy adopted in OED consists in relaxing the 0–1 constraints on the design variables and allowing them to additionally take any real values in the interval $[0, 1]$. In this way, the following much more convenient formulation is obtained:

Problem 2. Find $\mathbf{v}_{\text{D}}^* \in \mathcal{V} := \{\mathbf{v} \in [0, 1]^N : \mathbf{1}^\top \mathbf{v} = n\}$ to maximize $\Phi_{\text{D}}(\mathbf{M}(\mathbf{v}))$. Each minimizer \mathbf{v}_{D}^* is called a D-optimum *relaxed* design.

3 Sparsity Enforcement in Relaxed Designs

It goes without saying that the relaxed design \mathbf{v}_{D}^* may have many fractional components. On the one hand, there is an acute need to convert it to a design desirably with binary weights. Mathematically, the transformed design ought to be characterized by minimal support, i.e., a minimal number of nonzero components. On the other hand, however, the loss in the attained extreme value of the D-optimality criterion should be as small as possible.

A convenient device to quantify the closeness of any design $\mathbf{v} \in \mathcal{V}$ to \mathbf{v}_{D}^* in terms of the D-optimality criterion is the ratio

$$\mathcal{E}_{\text{D}}(\mathbf{v}) = \frac{\det^{1/m}(\mathbf{M}(\mathbf{v}))}{\det^{1/m}(\mathbf{M}(\mathbf{v}_{\text{D}}^*))}, \quad (5)$$

which is called D-efficiency, cf. [3]. A maximum value of this positive fraction is one, which is attained at any relaxed D-optimum design. Thus, the larger the efficiency, the better.

Writing $\|\mathbf{v}\|_0$ for the number of nonzero components of the vector \mathbf{v} , we introduce the following formulation to address the dilemma mentioned above:

Problem 3. Find $\mathbf{v}_{\text{D},0}^* \in \mathcal{V} := \{\mathbf{v} \in [0, 1]^N : \mathbf{1}^\top \mathbf{v} = n\}$ to minimize $\|\mathbf{v}\|_0$ subject to

$$\mathcal{E}_{\text{D}}(\mathbf{v}) \geq \eta, \quad (6)$$

where η signifies a given minimal acceptable value of D-efficiency. Each minimizer $\mathbf{v}_{\text{D},0}^*$ is called a minimum-support *relaxed* design with guaranteed D-efficiency η .

Thus, at the price of reducing some degree of D-optimality, which is controlled by η , we hope to find a sparse design as a recompense.

By abuse of notation, $\|\cdot\|_0$ is commonly called the ℓ_0 -norm, although it is not really a norm. Direct minimization of this function is problematic due to its discontinuous nature. A way to circumvent this complication is to approximate it by $\|\mathbf{v}\|_q = \left(\sum_{i=1}^N |v_i|^q\right)^{1/q}$ for a fixed small positive fractional q [17]. This function fails to be a norm, either, but it is much more convenient to use.

We employ this approximation and replace Problem 3 by the following ultimate formulation:

Problem 4. Find a vector $\mathbf{v}_{D,q}^* \in \mathbb{R}^N$ to minimize

$$J(\mathbf{v}) = \|\mathbf{v}\|_q^q = \sum_{j=1}^N w_j^q \quad (7)$$

for a fixed $q \in (0, 1)$ over the feasible set $\mathcal{W} := \{\mathbf{v} \in \mathcal{V} : \mathcal{E}_D(\mathbf{v}) \geq \eta\}$.

The concavity of the objective function (7) and the convexity of the admissible set \mathcal{W} make it a problem of concave programming [4]. Problems of this type are extremely difficult to solve. In fact, in contrast to convex programming problems, they usually contain many local minima. However, the compactness of the feasible set \mathcal{W} implies that Problem 4 must have a global minimum that is an extreme of \mathcal{W} .

3.1 Majorization-Minimization Algorithm

In recent years, numerous concave programming problems have been successfully solved using the majorization-minimization (MM) algorithm [18]. It turns out that this powerful computational scheme can be employed for the setting discussed here, too.

Starting from an arbitrary feasible initial point $\mathbf{v}^{(0)}$, the MM algorithm constructs a sequence of feasible vectors $\{\mathbf{v}^{(\kappa)}\}_{\kappa=0}^\infty$ by minimizing in each iteration a surrogate function $\mathbf{v} \mapsto \Psi(\mathbf{v}|\mathbf{v}^{(\kappa)})$. This function should be a convex tangent majorant of $J(\mathbf{v})$ at $\mathbf{v}^{(\kappa)}$, see [16] for details.

In view of the concavity of J , it is a simple matter to construct such a surrogate function here. Indeed, given $\mathbf{v}^{(\kappa)}$, we have $J(\mathbf{v}) \leq J(\mathbf{v}^{(\kappa)}) + (\mathbf{v} - \mathbf{v}^{(\kappa)})^\top \nabla J(\mathbf{v}^{(\kappa)})$, $\forall \mathbf{v} \in \mathcal{W}$, and the right-hand side of this inequality is convex in \mathbf{v} , overestimates $J(\mathbf{v})$ and is tangent to $J(\mathbf{v})$ at $\mathbf{v}^{(\kappa)}$ as the first-order Taylor approximation of $J(\mathbf{v})$. Therefore, we set

$$\Psi(\mathbf{v}|\mathbf{v}^{(\kappa)}) = J(\mathbf{v}^{(\kappa)}) + (\mathbf{v} - \mathbf{v}^{(\kappa)})^\top \nabla J(\mathbf{v}^{(\kappa)}). \quad (8)$$

The consecutive iterates of the MM algorithm are then defined as

$$\mathbf{v}^{(\kappa+1)} = \arg \min_{\mathbf{v} \in \mathcal{W}} \Psi(\mathbf{v}|\mathbf{v}^{(\kappa)}). \quad (9)$$

The sequence $\{J(\mathbf{v}^{(k)})\}$ is strictly decreasing until a minimizer is attained. What is more, any limit point \mathbf{v}^* of $\{\mathbf{v}^{(k)}\}$ is a stationary point of J [18]. Note, however, that the function J may have multiple local minima and the MM procedure usually terminates in one of them.

4 Minimization of the Surrogate Function via Generalized Simplicial Decomposition

The linearity of the surrogate objective function $\Psi(\cdot|\mathbf{v}^{(k)})$, the polyhedral form of the set \mathcal{V} and the convexity of the constraint (6) make the optimization problem (9) ideal for the use of generalized simplicial decomposition (GSD) to quickly solve it and drastically reduce the problem dimensionality, cf. [5] and applications in [20, 22]. In each iteration τ the dimensionality in GSD is reduced by inner approximation of the polyhedral set \mathcal{V} in \mathbb{R}^N by the convex hull of the set $\mathcal{V}^{(\tau)}$ of hitherto found extreme points of \mathcal{V} . Computations alternate between minimization of $\Psi(\mathbf{v}|\mathbf{v}^{(k)})$ over $\text{conv}(\mathcal{V}^{(\tau)})$, subject to the additional side constraint (6), and the search for a new extreme point $\mathbf{v}_{\text{xtrem}}^{(\tau)} \in \mathcal{V}$ to be included in $\mathcal{V}^{(\tau)}$. The former subproblem is called the restricted master problem, or RMP, and the latter subproblem is termed the column generation problem, or CGP. Substantial gains in the amount of computations are obtained as long as the cardinality of $\mathcal{V}^{(\tau)}$ is much lower than N , which is most often the case.

In the GSD, the solution of the RMP calls for a constrained-optimization method returning the Lagrange multipliers, but currently this is not a big issue since numerous solvers provide this option. In the CGP the gradients ∇J and $\nabla \mathcal{E}_{\mathcal{D}}$ evaluated at the solution of the RMP and the associated Lagrange multipliers are used to form a direction along which an extreme point of \mathcal{V} lying furthest from the current RMP solution is to be found. This constitutes a linear programming problem, which is easy to solve without resorting to the traditional simplex method [13, 22].

5 Illustrative Example

Consider the heat-conduction process through a thin isotropic square plate whose flat surfaces are insulated. Let $y(\mathbf{x}, t)$ be the temperature of the plate at spatial point $\mathbf{x} = (x_1, x_2) \in \Omega = (0, 1)^2$ and time instant $t \in T = (0, 1]$. Mathematically, the temperature distribution evolves according to the parabolic equation

$$\frac{\partial y}{\partial t} = \mu \Delta y, \quad (10)$$

where $\Delta y = \partial^2 y / \partial x_1^2 + \partial^2 y / \partial x_2^2$ and $\mu > 0$ stands for the diffusion coefficient. This equation is complemented with the boundary conditions

$$y(\mathbf{x}, t) = \begin{cases} cx_1 t & \text{on } (0, 1) \times \{0, 1\} \times T, \\ 0 & \text{on } \{0\} \times (0, 1) \times T, \\ ct & \text{on } \{1\} \times (0, 1) \times T, \end{cases} \quad (11)$$

where $c > 0$ defines the temperature increase rate on top, bottom and left boundaries, and the initial condition

$$y(\mathbf{x}, 0) = \alpha \sin(\pi x_1) \sin(\pi x_2) \quad \text{in } \Omega, \quad (12)$$

where $\alpha > 0$ signifies the maximum initial temperature.

The temperature evolution results from the decay of the initial state and the temperature changes forced at the boundaries. Setting $v(\mathbf{x}, t) = y(\mathbf{x}, t) - cx_1t$, we can easily deduce that v satisfies the equation $\partial v / \partial t = \mu \Delta v - cx_1$ subject to homogeneous Cauchy boundary conditions. Applying the Fourier method [7] to this equation, we can then express the solution of (10)–(12) in the following form:

$$y(\mathbf{x}, t) = cx_1t + 2 \sum_{k, \ell=1}^{\infty} y_{k\ell}(t) \sin(k\pi x_1) \sin(\ell\pi x_2), \quad (13)$$

where

$$y_{k\ell}(t) = -\frac{\gamma_{k\ell}}{\lambda_{k\ell}} + e^{\lambda_{k\ell}t} \left(y_{k\ell}^0 + \frac{\gamma_{k\ell}}{\lambda_{k\ell}} \right), \quad (14)$$

$$\lambda_{k\ell} = -\mu(k^2 + \ell^2)\pi^2, \quad (15)$$

$$\gamma_{k\ell} = \begin{cases} \frac{(-1)^k 4c}{k\ell\pi^2} & \text{if } \ell \text{ is odd,} \\ 0 & \text{if } \ell \text{ is even,} \end{cases} \quad (16)$$

$$y_{k\ell}^0 = \begin{cases} \frac{\alpha}{2} & \text{if } k = 1 \text{ and } \ell = 1, \\ 0 & \text{otherwise} \end{cases} \quad (17)$$

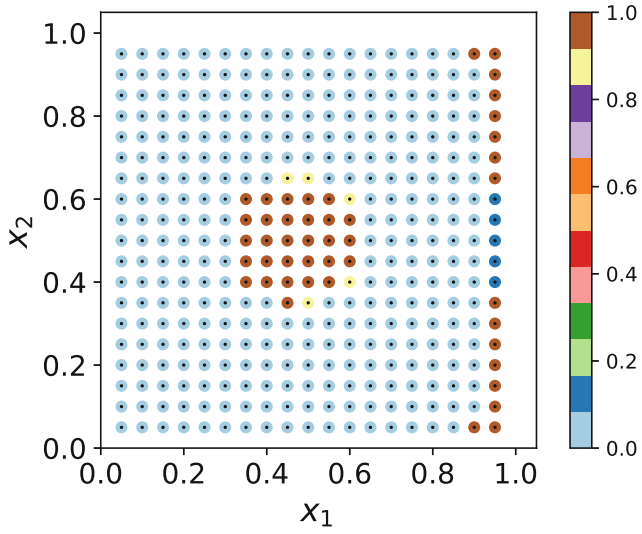
for $k, \ell \in \mathbb{N}$.

We treat coefficients μ , α and c as unknown parameters to be estimated based on measurements from temperature sensors to be deployed in Ω . Thus, $\boldsymbol{\theta} = (\mu, \alpha, c)$ and $\boldsymbol{\theta}^0 = (0.1, 1.0, 1.0)$ is used as the corresponding nominal value.

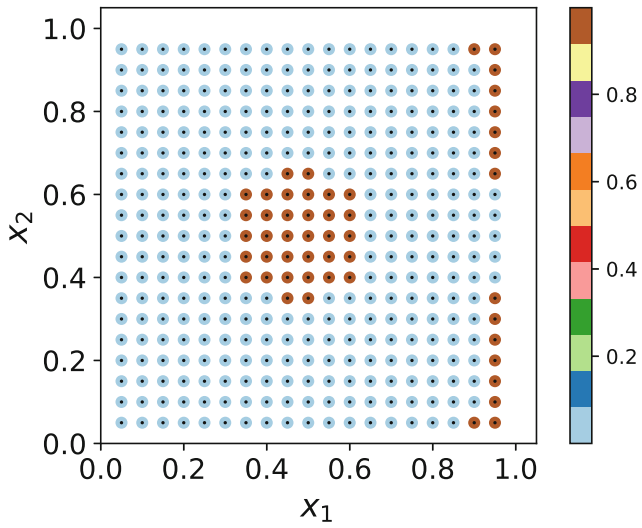
There are $n = 50$ sensors which can be located at gauged sites selected from a total of $N = 361$ candidate sites with coordinates $(i/20, j/30)$ for $i, j = 1, \dots, 19$. At each gauged site, the associated sensor is going to measure temperature at a sequence of time instants $t_k = 0.2k$, $k = 0, \dots, 5$.

All MM and GSD algorithms discussed in this paper were implemented in Python 3.9. In Eq. (13) the infinite sums were truncated (the infinity was replaced by a value of 10). The RMP problem in GSD was solved using the trust-constrained NLP solver (`scipy.optimize` with `method = 'trust-constr'`) available in SciPy. The code was run with the open-source Anaconda distribution 2022.10 under Windows 10 on a laptop equipped with Intel Core i7–6700HQ CPU, 2.60 GHz, 24 GB RAM.

Figure 1 shows the original relaxed D-optimum design weights and the same weights post-processed with the proposed algorithm for the minimal acceptable D-efficiency level set at 0.95 and $q = 0.2$. Only 17 iterations of the MM scheme were needed for convergence. Overall, the proposed postprocessing took no more than half a second.



(a)



(b)

Fig. 1. Optimum sensor configurations: relaxed D-optimum design with some nonnegligible weights in the centre and top-left subregions (a), sparse solution with a guaranteed D-efficiency of $\eta = 0.95$ (b). Black points denote candidate sites and the colours of the discs around them represent the weight values.

Observe that in the relaxed solution the nonzero weights cluster in the centre of the plate (this is where the maximum value of the initial state is located) and along the right boundary (this is where the maximum temperature increase on the boundary is exerted). Unfortunately, several weights are clearly fractional (the yellow points on the edges of the central cluster and the blue points in the middle of the right boundary). They are removed by the proposed MM scheme, which cleans the relaxed design and produces a design in which all weights are practically zero or one. Surprisingly, the constraint on the efficiency level turned out to be inactive, as the efficiency of the ultimate sparse solution was 0.998697. This means that in the close vicinity of relaxed designs there may be sparse solutions of a similar quality.

6 Conclusion

The key contribution of this article is an effective method of constructing solutions for relaxed formulations of optimal sensor location problems, characterized by components as close as possible to binary values within an acceptable deterioration of the design criterion value. Its advantage lies in solving a sequence of simple low-dimensional convex optimization problems, which are relatively easy to implement. The relaxed solutions so obtained can be easily employed to produce actual sensor configurations, which has been the biggest disadvantage of the relaxed approaches so far.

References

1. Alexanderian, A.: Optimal experimental design for infinite-dimensional Bayesian inverse problems governed by PDEs: a review. *Inverse Prob.* **37**(4), 043001 (2021)
2. Alexanderian, A., Petra, N., Stadler, G., Ghattas, O.: A-optimal design of experiments for infinite-dimensional Bayesian linear inverse problems with regularized ℓ_0 -sparsification. *SIAM J. Sci. Comput.* **36**(5), A2122–A2148 (2014). <https://doi.org/10.1137/130933381>
3. Atkinson, A.C., Donev, A.N., Tobias, R.D.: *Optimum Experimental Designs*, with SAS. Oxford University Press, Oxford (2007)
4. Benson, H.P.: Concave programming. In: Floudas, C.A., Pardalos, P.M. (eds.) *Encyclopedia of Optimization*, pp. 462–466. Springer, Boston (2009). https://doi.org/10.1007/0-306-48332-7_68
5. Bertsekas, D., Yu, H.: A unifying polyhedral approximation framework for convex optimization. *SIAM J. Optim.* **21**(1), 333–360 (2011). <https://doi.org/10.1137/090772204>
6. Chepuri, S.P., Leus, G.: Sparsity-promoting sensor selection for non-linear measurement models. *IEEE Trans. Sig. Process.* **63**(3), 684–698 (2015)
7. Dautray, R., Lions, J.L.: *Mathematical Analysis and Numerical Methods for Science and Technology: Evolution Problems I*, vol. 5. Springer, Berlin (2000)
8. Gejadze, I.Y., Shutyaev, V.: On computation of the design function gradient for the sensor-location problem in variational data assimilation. *SIAM J. Sci. Comput.* **34**(2), B127–B147 (2012). <https://doi.org/10.1137/110825121>

9. Haber, E., Horesh, L., Tenorio, L.: Numerical methods for the design of large-scale nonlinear discrete ill-posed inverse problems. *Inverse Prob.* **26**(2), 025002 (2010)
10. Herzog, R., Riedel, I., Uciński, D.: Optimal sensor placement for joint parameter and state estimation problems in large-scale dynamical systems with applications to thermo-mechanics. *Optim. Eng.* **19**(3), 591–627 (2018). <https://doi.org/10.1007/s11081-018-9391-8>
11. Joshi, S., Boyd, S.: Sensor selection via convex optimization. *IEEE Trans. Sig. Process.* **57**(2), 451–462 (2009)
12. Patan, M.: *Optimal Sensor Networks Scheduling in Identification of Distributed Parameter Systems*. LNCIS, Springer, Berlin (2012). <https://doi.org/10.1007/978-3-642-28230-0>
13. Patan, M., Uciński, D.: Generalized simplicial decomposition for optimal sensor selection in parameter estimation of spatiotemporal processes. In: 2019 American Control Conference (ACC), pp. 2546–2551 (2019). <https://doi.org/10.23919/ACC.2019.8815091>
14. Pronzato, L., Pàzman, A.: *Design of Experiments in Nonlinear Models. Asymptotic Normality, Optimality Criteria and Small-Sample Properties*. LNS. Springer, New York (2013). <https://doi.org/10.1007/978-1-4614-6363-4>
15. Rafajłowicz, E.: Optimal input signals for parameter estimation: In *Linear Systems with Spatio-Temporal Dynamics*. De Gruyter, Berlin (2022)
16. Scutari, G., Sun, Y.: Parallel and distributed successive convex approximation methods for big-data optimization. In: Facchinei, F., Pang, J.-S. (eds.) *Multi-agent Optimization*. LNM, vol. 2224, pp. 141–308. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-97142-1_3
17. Sun, X., Zheng, X., Li, D.: Recent advances in mathematical programming with semi-continuous variables and cardinality constraint. *J. Oper. Res. Soc. China* **1**(1), 55–77 (2013). <https://doi.org/10.1007/s40305-013-0004-0>
18. Sun, Y., Babu, P., Palomar, D.P.: Majorization-minimization algorithms in signal processing, communications, and machine learning. *IEEE Trans. Sig. Process.* **65**(3), 794–816 (2017)
19. Uciński, D.: *Optimal Measurement Methods for Distributed-Parameter System Identification*. CRC Press, Boca Raton (2005)
20. Uciński, D.: Construction of constrained experimental designs on finite spaces for a modified E_k -optimality criterion. *Int. J. Appl. Math. Comput. Sci.* **30**(4), 659–677 (2020). <https://doi.org/10.34768/amcs-2020-0049>
21. Uciński, D.: D-optimal sensor selection in the presence of correlated measurement noise. *Measurement* **164**, 107873 (2020). <https://doi.org/10.1016/j.measurement.2020.107873>
22. Uciński, D.: E-optimum sensor selection for estimation of subsets of parameters. *Measurement* **187**, 110286 (2022). <https://doi.org/10.1016/j.measurement.2021.110286>
23. Uciński, D., Patan, M.: D-optimal design of a monitoring network for parameter estimation of distributed systems. *J. Global Optim.* **39**(2), 291–322 (2007)



Earned Value Method in Public Project Monitoring

V. M. Molokanova^(✉)

Dr. Sc. (Tech.), National Technical University “Dnipro Polytechnic”, Dnipro, Ukraine
molokany@gmail.com

Abstract. The article provides an overview of project management development based on the earned value method at the present time. A separate discussion of the main factors related to the use of the earned value analysis in the management of public projects. The article considers the vision in the project management model of two issues: planning models and development models based on factual results. It gives the opportunity to show the project state at an early stage, and explain the core rules for possible management. The methodology of the implementation of an innovative project’s portfolio was initiated based on the analysis of actual demonstrations of the portfolio components, the earned value method, and forecasting costs for existing project groups. The proposed methodology allowed us to obtain good results for developing public project management. Approbation was carried out at the enterprise of LLC “CI-Center” (Dnipro, Ukraine). The obtained results allow for strengthening the method of informing about the introduction of the portfolio and constant adjustment based on the choice of possible ways of further development. A dynamic model of project management, planning, monitoring, and implementation of projects has been proposed for use.

Keywords: Earned Value Method · Performance Indicators · Public Project Monitoring

1 Introduction

The project is, first of all, a unique system of planned changes, and as any uniqueness means uncertainty, both the flow of the process itself and the expected results. One of the important tasks of the project manager is to track the progress of work in the project. However, it is impossible to be sure in advance that the plan for creating a unique product can take into account all aspects and implementation [1]. The skills of the project manager are manifested through the ability to identify the necessary improvements in a timely manner and make adjustments to bring the project to successful implementation. It is the monitoring results that are the basis for adjusting the decisions made earlier, if the deviations during the project implementation are significant. During the implementation of the project, managers should in a short time not only assess the deviations that have arisen in the state and the project, but also calculate their impact on the main indicators of project success and find timely decisions on changes in the project implementation.

2 Literature Review

In the methodology of project management, the first net models and Program Evaluation Review Technique appeared in the 50s of the 20th centuries, at the same time they began to determine the optimal end date by the method of critical path [2]. In 1962, the PERT/Cost technique was introduced, which took into account not only time, but also cost characteristics. In the U.S. Department of Defense's project standards system, the Cost/Schedule Control Systems Criteria technique has been mandatory since 1967, but it is considered rather cumbersome and complex. Its complexity leads to the fact that in commercial projects less and less used C/SCSC analysis [3]. However, Harold Kerzner [4] considers C/SCSC analysis a relevant organization maturity in project management. In the 90s of the 20th centuries, a simplified method of mastered volume, provided in most textbooks, was becoming more and more widespread [5]. At the same time, more and more attention paid to the quality of project management, and international standards ISO 10006:1997, ISO 10007:1995, ISO 9000:2000 appear, which are adopted in a number of countries as national standards.

It should be noted that in recent years there has been a significant number of works and manuals on project management, and in general, in the methodological literature, according to the author, insufficient attention has been paid to the issues of managing the implementation of the project and automating the processing of the results of project monitoring. Only a few authors pay enough attention to these processes [6]. As a rule, this issue is more discussed in the section "Managing the cost of the project", but does not have a methodology and mathematical algorithms for making management decisions in the process of monitoring the project. The most adapted to practical use are graphic tools using the method of mastered volume, which have appeared in recent years [7]. Most authors when using the earned value method make assumptions about stabilizing changes in indicators during the project, while in reality it may be quite the opposite. Statistics indicate that the number of deviations in the project increases as it is implemented. Detailed studies of military-industrial projects in the United States have shown an increase in current overspending on the completion of the project [8]. Unfortunately, in Ukraine there are no statistics on the dynamics of the main indicators of earned value method in public projects, so the task of creating archives of implemented projects remains relevant.

3 Unsolved Aspects of the Problem

Analysis of the difficulties identified on many projects shows that the experience and intuition of managers are not always able to ensure that the right decision is made in difficult conditions. On the other hand, changes in the project can be depicted in a fairly strict mathematical form, that is, formalized. This means that situations arising in the practical implementation of the project can be simulated, and options for the most appropriate solutions for their management can be obtained from the analysis of modeling results. To obtain reasonable rational management solutions, science has accumulated a sufficient number of complex methods and models that the project manager needs to consider in a short time. The modern level of calculated equipment and means of information

transmission allows you to automate many stages of collection, planning and processing of information on changes in the state of the project, predict their further development, determine their impact on the main indicators of project success – completion date, cost, quality, calculate or simulate various solutions to overcome deviations from the project plan that have arisen, determine the most appropriate measures, ensuring the permissible effectiveness of the project. In public projects, it is especially important to show the progress of the project [9]. However, due to the significant uncertainty of such projects, their transparent tracking becomes quite a difficult process.

The purpose of this article is to study the state of knowledge on managing the implementation of projects based on the earned value method at the present time. It is proposed to distinguish two components in the project implementation management model: a planning model and a decision-making model based on the results of observations. A single dynamic project management unit is proposed, combining the processes of planning, observing, redeveloping and implementing changes.

4 Method Description

The research methodology provided for a descriptive analysis of the real case under study, the implementation of a number of public projects for the reconstruction of the city of Dnipro. The research methodology, in this case, includes a simple descriptive real study of the author on the application of the method of mastering volume. Reports from projects were used to obtain data for the study. As a simple descriptive example of research, logic has been used to test the possibility of repeating the earned value analysis process in public projects. The developed methodology allowed to obtain certain results of approbation for monitoring the management of public projects. The approbation was carried out at the enterprise LLC “SI-Center” (Dnipro, Ukraine). The results suggest that the proposed method is effective in obtaining information on the progress of the implementation of the development project portfolio and constant adjustment it to the choice of possible ways of further development. All data, reports, databases, perceptions and concerns were received from project participants. They were the main source of information in this study in addition to the reports. The earned value analysis method is already considered as a relevant community of project managers and cost engineers. The theoretical references related to earned value analysis have been investigated with the aim of comparing with practical results and guaranteeing the value of this real case study through real results. The descriptive case was supplemented by research, since the author took an active part with the project team.

5 Presentation of Results

The implementation of the project is among the phases of planning. Planning processes are among the most important project processes, because the result of planning is a unique product, which is any project [2]. Planning helps to transform the vision of the project by the customer, described in qualitative categories, into a structured model, that is, to translate a complex, poorly marked task into a problem that has a solution defined in quantitative indicators. The essence of planning consists in fixing ways to achieve

certain goals by formation of a complex of processes to be implemented. Planning can begin with any document that defines the requirements and expectations of the customer regarding the project, for example, a statement of work (Scope of Work) or a memorandum of understanding (Memorandum of Work). Planning covers all stages of the creation and implementation of the project [10], starting with the development of the project concept, continues when choosing strategic solutions of the project and then goes into a breakdown into works. As a result of planning, we are gradually moving from a complex unstructured, problem to managing the quantitative and qualitative indicators of the project.

Planning processes are carried out taking into account the specifics of the project, its type, scale, timing. Plans, graphs, networks, as a reflection of planning processes, form a certain hierarchical structure. Breaking down plans at the level is an effective tool that allows you to manage complex projects. The plan's aggregation levels must match the management levels. The higher the level, the more aggregated, generalized information should be. According to the levels of planning can be distinguished: conceptual planning, strategic planning, detailed planning. Conceptual planning is carried out at the stage of project initiation and can take place without frequent project manager, but may contain an enlarged calendar plan, financial plan, documentation on control and responsibility [11]. Strategic planning is the process of developing enlarged strategic plans that align the project strategy with the overall strategy of the enterprise. Detailed planning is associated with the development of schedules for operational management at the level of responsible performers. The level of complexity of the project implementation managing depends on the content of the project, its innovation, repeatability and many other factors. In portfolios of public projects, we can count four types of projects having different degrees of uncertainty risks when planning by the method of critical path:

1. Projects that are an indefinite set of operations and are more suitable for the definition of "innovation processes".
2. Projects that are constantly repeated in a certain way and can be defined as "uncomplicated".
3. Projects are large in time and complex, uncertain in beginning and end, which can be called as "programs".
4. Projects very short, simple that almost do not require managerial efforts.

Possible expansion of the element base of the classical methodology of project management is shown on Fig. 1. Of course, this interpretation of the typology of projects will rub the excellent practice of planning and determining the deadline of the project, which leaves a wide field of research to complement the methodology of project management by expanding the element base.

The main indicators that characterize the progress of work on the implementation of the project and are used in the earned value method [5]:

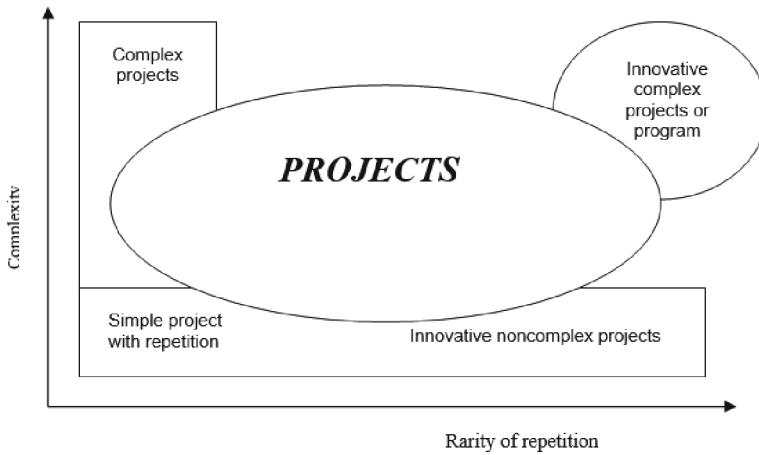


Fig. 1. Expansion of the elements of project management

BCWS - the budget cost of the planned works;

ACWP - the actual cost of the works performed on the date of inspection;

BCWP – the budget cost of works performed on the date of inspection;

CV = BCWP - ACWP – cost deviation;

SV = BCWP - BCWS – planned deviation;

BCWP/ACWP = CPI - cost performed index;

BCWP/BCWS = SPI – schedule performed index;

ETS – general estimate to completion;

EAC – general estimate for completion.

The literature proposes the use of the earned value method to evaluate the project upon completion. Evaluation upon completion is a forecast of the amount of costs or deadlines that should be expected upon completion of all design work. It is clear that the cost of the project upon completion is calculated as: $EAC = ACWP + ETC$.

According to the method of mastered volume, it is possible to calculate an optimistic and pessimistic forecast of the cost of the project upon completion:

Optimistic forecast

$$EAC = (BC - BCWP)/CPI + ACWP; \quad (1)$$

Pessimistic forecast

$$EAC = (BC - BCWP)/CPI * SPI + ACWP; \quad (2)$$

The current control of the state of the project uses calculations of the cost development index CPI and the index of development of the volume SPI.

If, CPI - the cost development index is less than one, this means a lag behind the plan in terms of each hryvnia spent from the budget, if the SRI is more than one, the progress of work goes with cost savings.

If the indicator SPI - the volume development index is less than one, for example, it is equal to 0.66, this means that at the control date the work was completed by 0.66 UAH. For each hryvnia according to the calendar schedule. That is, if the earned volume index is less than one, this means a lag behind the timing of the project (Table 1).

Table 1. Project development indexes

Value	Cost Performed Index	Schedule Performed Index
>1	Below cost	Ahead of schedule
=1	Corresponds to the cost	Coincides with the schedule
<1	Above cost	Lagging behind schedule

But the right is that the use of estimates (1), (2) is valid only if there is a constant linear relationship between costs and the amount of work with fixed resources. For the general case of the relationship between the volume of work performed and the costs, such a directly proportional dependence can only be considered as an assumption. That is, for the general case, it is necessary to assess the dependence of actual costs on time on the basis of observation of actual and planned costs and describe it using the existing mathematical apparatus.

The main goal of managing the implementation of the project according to the earned value method is the possibility of early detection of the discrepancy between the actual indicators and the planned ones and forecasting the future results of the project – the timing and costs at the end. To predict the results of the project, it is proposed to use [8] following the derived indicators of the mastered volume:

- $\Delta_0(t) = c_0(t) - c(t)$ - the difference between planned and actual costs;
- $\Delta(t) = C_0(t) - c_e(t)$ - the difference between planned and mastered costs;
- $\Delta_e(t) = c(t) - c_e(t)$ - the difference between actual and mastered costs;
- $\alpha(t) = c_e(t)/c_0(t)$ - earned value indicator (SPI);
- $\beta(t) = c_e(t)/c(t)$ - cost development indicator (CPI).

The value of the actual costs of the project is the main indicator, which is estimated during the implementation of the project. Since the indicator $\beta(t)$ reflects the efficiency of the use of funds, then at the time of t the value of C (total costs for the project) can be obtained as the sum of the funds already spent and the funds remaining until the completion of the project. The latter value is defined as the proportion of the difference between the planned value of the total costs and the mastered amount of funds to the efficiency of the use of funds:

$$C(t) = c(t) + (C_0 - c_e(t))/\beta(t) \tag{3}$$

It is also possible to use the “pessimistic” forecast of estimating the total costs of the project (PMI, 2013), where the efficiency of the use of funds for the project is calculated

as $\alpha(t) \cdot \beta(t)$:

$$C(t) = c(t) + (C_0 - c_e(t))/\alpha(t) \cdot \beta(t); \tag{4}$$

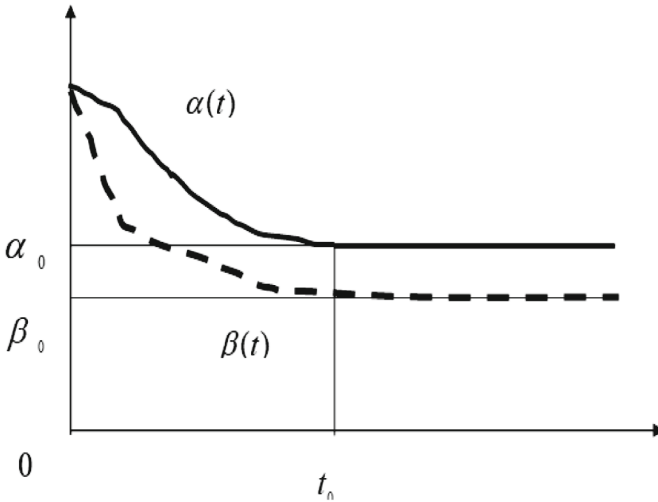


Fig. 2. Stabilization of efficiency coefficients.

It is clear that such an assessment suggests that over time $\alpha(t)$ and $\beta(t)$ remain unchanged, that is, stabilized. Today, most of the known tools for using the earned value method are based on the assumption of stabilization of these indicators and during the implementation of the project (Fig. 2).

Thus, using the predictive formulas (1) and (2), you can write the completion time of the project as:

$$T = T_0/\alpha_x(t); \tag{5}$$

And the actual costs of the project:

$$C = X_0/\beta(t); \tag{6}$$

It is clear that the use of the method would have good consequences, if the project consisted of one work. Of course, in practice there are no such projects. Therefore, we have to consider a project that has a significant number of works of different levels, based on well-known project models - the structure of the breakdown of works, the network model and the Gantt schedule.

Thus, in order to apply the earned value method for the entire project, it is necessary to consider the planned indicators, actual indicators and indicators of the mastered volume for each work, measured in costs. But in the project, you need to have indicators that reflect the results of work in physical units (for example, length, pieces, area, etc.), and not monetary units. In addition, for management purposes for all work, it is necessary

to measure not only the value of indicators, but also to be able to monitor the dynamics of changes in these indicators. Therefore, in some studies [1, 2, 8] it is proposed to introduce three more indicators of the "physical" volume – then just the planned, actual and mastered volume.

Then each work in the project can be described by six derivatives:

$C_0(t)$ – Budget Cost of Work Scheduled (BCWS);

$C(t)$ – Actual Cost of Work Performed (ACWP);

$C_e(t)$ – Budget Cost of Work Performed (BCWP);

$x_0(t)$ – Budget Quantity of Work Scheduled (BQWS);

$x(t)$ – Actual Quantity of Work Performed (AQWP);

$x_e(t)$ – Budget Quantity of Work Performed (BQWP).

The project will be completed when the mastered volume coincides with the total amount of work on the project, since the amount of work performed is an indicator of project execution, not costs. At the same time, the deadline and costs are indicators of the effectiveness of the project, but in the performance indicators there is a ratio of the mastered volume not to the mastered costs, but to the actual one.

The proposed indicators of mastered volume are not independent – as a rule, there is a technological relationship between costs (resources) and volume. If you depict this relationship as an operator $G_0(\cdot)$, then there is $x_0(t) = G_0(c_0(t))$. If for external reasons it is possible that $c(t) \neq c_0(t)$, then this will cause a difference in the actual volume and planned. In addition, due to internal reasons (for example, planning errors), an incorrect assessment of the operator is possible. $G_0(\cdot)$ - In fact, there is a relationship between cost and volume $x(t) = G(c(t))$. Then the discrepancy between the actual and mastered volume will cause an error $G(\cdot) \neq G_0(\cdot)$. Thus, for effective management of the project implementation, it is necessary to take into account not only external, but also internal reasons for the discrepancies between the planned and actual indicators. Having solved these two problems, you can begin to predict future indicators of the mastered volume. It is clear that to predict the results of the project on the basis of its current observations, it is possible to apply mathematical forecasting methods (least squares method, exponential smoothing method) and the method of expert assessments. When predicting a certain indicator, it is assumed that its future value somehow depends on past periods. Analyzing the past, it is necessary to determine the average rate of change in the indicator for its future assessment. If there is a proportional dependence, the future is easy to determine. If there are sharply different assessments, then smoothing and additional analysis are required. Choosing an anti-aliasing method should ensure that the trend of changing the indicator persists.

For example, if you define an indicator as a variable $\alpha(t_i)$, then according to the moving average method, the smoothed variable is the average value of several adjacent values, for example:

$$\alpha_i = \frac{1}{3}(\alpha_{i-1} + \alpha_i + \alpha_{i+1}) \quad (7)$$

If, for example, the last period is the most important, it can be added a coefficient of "weight". Since the choice of importance is always subjective, the project manager has the opportunity at some stage to "add weight" to the work of the project, if in his opinion they are decisive. Conversely, if the manager recognizes the unrepresentativeness of

the previous work for the project, then he can reduce its value by assigning it a lower coefficient in the calculations, that is, to calculate forecasting indicators, use the formula:

$$C = p_1 C_1 + p_2 C_2 + p_3 C_3 + \dots + p_m C_m; \quad (8)$$

where C_i , $i = 1, 2, 3 \dots m$ – actual costs of the work project, a P_i – is their weight coefficients. It is clear that to assign weight coefficients, we can use the opinions of experts and project participants.

The general algorithm for the applied use of the earned value method at the stages of planning, control and operational management can be determined as follows:

1. Determination of the scope of work according to the structure of the project decomposition.
2. Planning at the level of individual works using a software product.
3. Distribution of responsibility by levels of work in the project.
4. Development of a directive work schedule using a software product.
5. Evaluation of the actual implementation of the project and comparison with the directive schedule (indicators $\Delta(t)$ and $\alpha(t)$).
6. Cost effectiveness appraisal (indicators $\Delta_e(t)$ and s).
7. Forecasting the total actual costs of the project according to the observation of the progress of its realization using software.
8. Making a management decision on the results of forecasting.
9. Management of the remaining part of the project.
10. Checking the quality of work.

During the implementation of the project, when the manager has limited time and limited information and you need to make decisions in real time, forecasting software should be useful to minimize decision-making time. Moreover, the existing models of network planning today are already quite complex and require a large amount of information and time. So, to create identification methods, forecasting and operational management requires the creation of appropriate tools and their integration into existing project management software.

Thus, to solve the problem of forecasting within the framework of the method of mastered volume, a method of adapting the management of work on the project by smoothing the functions of deviations $C(t)$ and $C(t)$ by the method of moving average and coefficients of “weight” is proposed. The relevant rules for predicting future values of critical indicators of the project can be used (that is, to be “sewn up”) in finished software products. It is much better when the manager himself is able to integrate them into the corresponding computer programs, independently determining the need for smoothing for a particular project or part of the project. In this case, the calculations of the projects were performed in the program MS Office Excel and integrated not into MS Project.

After the completion of the formation of the portfolio of orders, the stage of its implementation begins. The earned value method allows you to track the status of the project portfolio at the beginning of its implementation. The basic schedule of project portfolio implementation, agreed before the start of its project implementation, will be carried out as efficiently as portfolio managers track and respond to its changes as they progress through time.

When evaluating actual and planned indicators, portfolio managers can monitor its implementation from the beginning of implementation to completion (although the development portfolio can continue indefinitely). The advantage of the earned value method is that if the portfolio manager receives unsatisfactory actual indicators at the beginning of its implementation, this may be a signal of unsuccessful planning and a reason to take actions to prevent undesirable consequences.

The general algorithm for managing the implementation of a portfolio of projects can be divided into four stages:

1. Collecting information, calculating deviations of actual indicators from planned ones and assessing the general condition of the portfolio.
2. Determining the causes of deviations of individual components of the portfolio according to the method of mastered volume.
3. Forecasting the future value of portfolio components by groups.
4. Making management decisions about influence on project.

In case of reformatting the portfolio, you need to return to the processes of its formation, optimization and approval of the basic plan of the portfolio. Thus, the processes of managing the implementation of the project portfolio are cyclical in nature and can be continued indefinitely. The cyclical nature of the implementation of the project portfolio constantly necessitates additional analysis and return to the already approved processes.

The collection of actual information on the dates of start and completion of work and the percentage of completion, the amount of resource use and costs allows us to use the earned value method to assess both the state of individual components of the portfolio and its total as of the date of verification. At the enterprise LLC "SI-Center" (Dnipro, Ukraine), a modified method of percentage of execution was used as an indicator for measuring the mastered volume.

All projects of portfolio are conditionally divided into time periods lasting on a working Sunday. To determine the percentage of earned volume for a separate component of the portfolio from the manager of this project, we have to get one of three answers about the status of works implementation: 0%, 50% or 100%. Such a not entirely accurate method is inexpensive and effective, even if there are some inaccuracies in the information received.

The emergence of a problem situation in a separate project may be associated with the following two types of reasons:

1. The general financial condition of the enterprise differs from the planned one, which means a lack of funding for projects (exogenous reasons).
2. Disadvantages of project planning and significant deviations of actual indicators from planned ones (endogenous reasons).

The solution of both types of problems is significantly influenced by the ability of project managers to predict the development of the situation, track structural connections and interpret the information received in the context of the general state of the existing enterprise. It is generally known that it is impossible to compensate for the costs.

Therefore, at the stages of portfolio implementation, changes are constantly made to the basic plan of portfolio costs and resources are redistributed among the components

of the portfolio. The algorithm of the methodology for managing the project portfolio at the stage of its implementation is provided in Table 2.

Table 2. Algorithm for managing a portfolio of projects at the stage of implementation

№	Stage	The content of the work performed
1	Documenting actual project indicators as of the date of verification	Form of information on portfolio components as of the date of translation
2	Using the metrics aggregation tool for your entire portfolio	Form of information on the portfolio implementation on the date of translation
3	Analysis of the state of the portfolio and forecasting of its indicators	Formation of a report on the implementation and adjustment of planned performance indicators
4	Tracking trends and frustration signals in the portfolio	Comparison of indicators for the current and past periods. Analysis of index indicators and CPI and SPI trends
5	Formation of proposals for changes	Return to portfolio formation processes
6	Making decisions on changes	Approve or reject changes
7	Return to p.1–6 on a new verification date	Formation of a report on on the new date of verification

Table 3. Summary data for the first eight weeks of project portfolio implementation

№ period	Planned Costs Thousand UAH	Actual Costs Thousand UAH	Resource utilization percentage %	Earned Volume Thousand UAH	Cost Performed Index	Schedule Performed Index
1	8000	8200	75	6000	0,73	0,75
2	11500	9300	60	12150	1,31	1,06
3	23500	21300	90	16650	0,78	0,71
4	24100	21900	83,3	17150	0,78	0,71
5	29500	27300	76,8	22656	0,83	0,77
6	39700	28200	90	44111,1	1,56	1,11
7	54800	46200	85	46580	1,01	0,85
8	72500	60500	90	65250	1,08	0,90

This algorithm allows you to determine the internal dynamics of the project portfolio at the very beginning of its implementation and creates the basis for making effective decisions. The possibility of intensive management using well-known software is considered on the basis of portfolio indicators, which are collected data for the first eight weeks of its implementation (Table 3).

Calculations of forecast data for indices of cost development and mastered volume are provided in Tables 4 and 5. The initial data correspond to the corresponding graphs (Figs. 3 and 4). A moving average chart shows the convergence of smoothed and actual data values. Smoothed forecasts of indices show trends in their changes.

Table 4. Calculation of the costs forecast

CPI	Forecast
0,73	#Н/Д
1,31	#Н/Д
0,78	0,94
0,78	0,96
0,83	0,80
1,56	1,06
1,01	1,13
1,08	1,22
	1,05
	1,08

Table 5. Calculation of the forecast by volume of work

SPI	Forecast
0,75	#Н/Д
1,06	#Н/Д
0,71	0,84
0,71	0,83
0,77	0,73
1,11	0,86
0,85	0,91
0,90	0,95
	0,88
	0,90

In Figs. 3 and 4 provided schedules of deviations in cost and scope of work, as well as forecasts for the first eight periods of the portfolio after the measures taken. The smoothed graph by volume shows that for the last period the mastered volume is slightly less than planned, but the index approaches one, that is, the position will be further leveled, and the situation does not need to be adjusted.

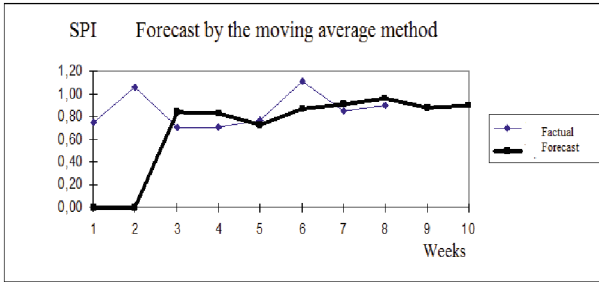


Fig. 3. Charts and forecasts of deviations in costs

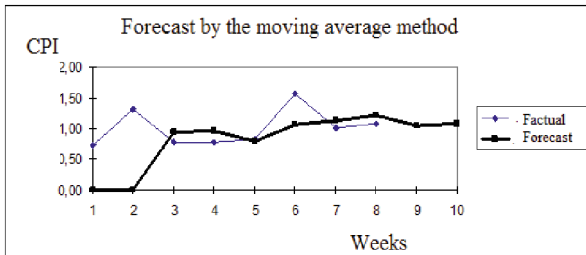


Fig. 4. Charts and forecast deviations by volume

The smoothed cost schedule shows that the development of costs slightly exaggerates the plan, but this is associated with an increase in the pace of development of the volume of work. The situation requires further monitoring so that a further increase in project costs (while maintaining the trend) does not exaggerate the planned reserve (20%).

An important element of the overall management of the development portfolio is the tracking of trends in the index indicators of the portfolio (see Table 3) which needs further explanation. As is known, the values of some economic indicators [54–56], which are random in nature, can be interpreted in the form of time series – the obtained data sequences at a certain point in time, where and – the ordinal number of the value of the empirical series on the time axis. Each such series is characterized by a certain trend in the development of processes called a trend. The forecast models of the mastered volume method offered, as defined in Sect. 3, provide for a functional relationship between costs and scope of work. Thus, to analyze the further development of the project portfolio, it is considered useful to analyze trends in deviations in time and costs. Trend models provide forecasts for short and medium periods under the following conditions:

1. The period of time analyzed must be sufficient to identify the pattern.
2. The processes described by the time series must have some inertia.
3. The autocorrelation function of the time series must be fading, that is, the influence of later information grows stronger over time (Figs. 5 and 6).

We assume that the project development processes meet all the general requirements of the trend model. The MS Excel software tool builds trending models graphically based

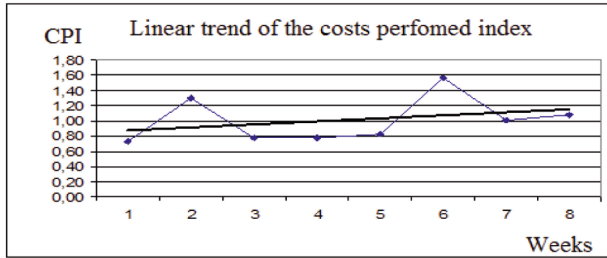


Fig. 5. Index of mastered costs with linear trend

on diagrams representing different dynamics. For an empirical data series, a diagram of the selected trend type is constructed according to a certain algorithm.

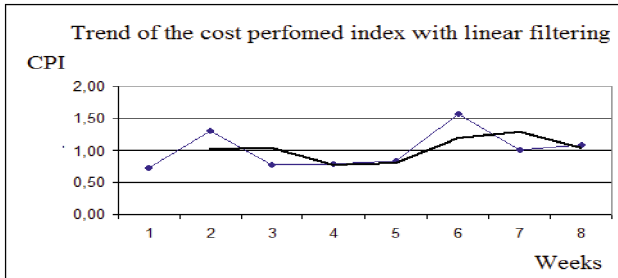


Fig. 6. Earned cost index and linear filtration trend

The diagram is transferred to the editing mode, a series is allocated to build a trend line, the commands “Chart - Trend Line” are executed, a window appears to select the type of trend line. The type of trend is selected taking into account the form of the series and the value of the approximation coefficient of reliability. For our case of quantities, according to the author, linear, logarithmic and polynomial approximations are most suitable.

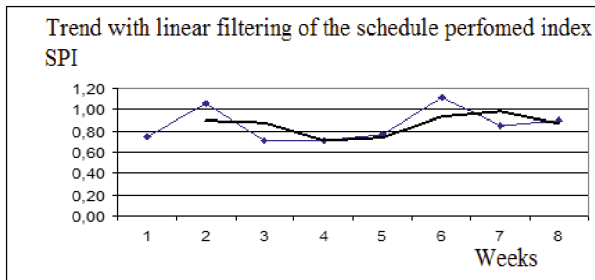


Fig. 7. Earned value index with linear trend

Using the table of summary data for the first eight periods of the project and the row of the earned cost index (see Table 3), a linear trend of the project cost development index and a trend with linear filtration are constructed (Fig. 7). This allows you to identify the main trend in the development of the project in terms of costs. As can be seen from the pic. 7–8 different types of trends show us that in the last period, the state of the portfolio changes in a good direction, the earned spending index approaches one. That is, the portfolio manager must make efforts to maintain the current trend.

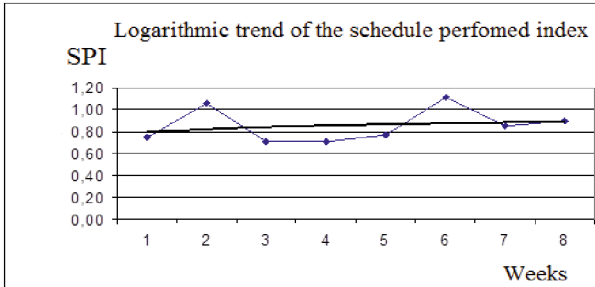


Fig. 8. Earned value index with logarithmic

The trends shown in Figs. 7 and 8 show that in general, the situation on the mastered volume of the project portfolio is changing in a good direction, the volume of work performed is increasingly approaching the planned one, and the index of the mastered volume to one.

Thus, reformatting the project portfolio in the process of monitoring it can be carried out by one of the following methods:

1. Enumeration according to the method of mastered volume.
2. Transfers using amendments to the indicators of cost development and execution of the schedule.
3. Enumeration using coefficients obtained by one of the empirical methods.
4. Setting new deadlines for the completion of individual projects at the discretion of managers based on trends.

Calculations of individual forecasts provide a real picture of the state of the project and allow the manager to realistically evaluate the results obtained, which is very important for intensive management of an innovative project. The ability to quickly predict allows the manager to better navigate the changes in the state of the project and deviations from the planned indicators and quickly resort to certain corrective actions. The sooner the corrective actions are taken, the better for the final results of the project.

In addition to determining the projected cost, the earned value method allows you to make forecasts of other indicators of the project. It also effectively allows you to determine the deviations of the project by time. Given the overall complexity of the dynamic management of project indicators as it is implemented, it is possible to predict further improvement in the use of earned volume indicators to make informed decisions, especially for work with a large innovative component. Thus, the proposed methodology

can significantly improve the management of development portfolio projects at the stage of their implementation.

At the present stage of development of information technology, it is impossible to practice project management without the use of computer technology. Thanks to the development of software and hardware, it became possible to accumulate large amounts of information, many intelligent systems, software products to support decision-making. This has led to the creation of new methods and means of intellectual processing of information, allowing the manager to more appropriately use his knowledge of the subject area. In turn, in project-oriented companies, the development of integrated intelligent project management systems that would combine both the existing methods of the subject area of project management and the provision of the processes of accumulation of internal organizational knowledge is becoming relevant. Achievements in the field of computer technologies allowed to automate corporate project management at the enterprise of LLC “SI-Center” (Dnipro, Ukraine) and create automated personal workplaces for project managers and ordinary users.

Project managers and members of the development portfolio management team exchange data regularly, as each manager sends their additions to the database (taking into account the rights to change), and they are combined on the main server. Every time a portfolio manager receives a request for changes, he uses one method or another to analyze the status of the entire portfolio on the active date. Planning specialists upload innovative projects to the main server and calculate the expiration dates of portfolio components. They can also align resources and generate reports on the current date, check the mutual impact of portfolio components, and return some reports to those responsible for revision. Thus, dynamic management of the portfolio of development projects combines the efficiency of using portfolio resources and the compliance of the portfolio with sustainable development priorities.

6 Conclusion

It was demonstrated that the concept of earned value is an effective technique in project management. The use of the proposed method allows us to draw up a detailed project schedule, depending on the complexity and innovation at the initial stage. This allows project managers to receive early warning signals to change the direction of the project.

EVM allows us to make better and more effective management decisions, minimizing adverse consequences for the project. Among the main advantages of using the earned values method:

1. Accuracy in achieving reporting.
2. Early warning, which provides a tool to project managers, allowing them to take the necessary corrective action if the project spends more money than was planned.

The effectiveness of the cost control system and schedules increases due to the calculation of trends in projects. They indicate how the project is developing and what the situation is in terms of the results obtained in the work schedule.

Thus, it has been demonstrated that earned value management allows us to eliminate the complexities of project planning by taking into account the risk of inaccurate planning.

It should also be noted that using the earned value method in public projects, we face difficulties in measuring added value of an intangible nature. The only recommendation here is the expert evaluation method. But the study of this topic will be continued in the next article.

If within the framework of the constructed model, the project manager has all the necessary indicators of the performed volume, then some of them can be considered managerial, for example, resource intensity, and the rest are dependent on them. That is, depending on the model chosen, some of the indicators can be managerial, and the other dependent is the one that is managed. At the same time, to solve the appropriate optimization problems, it is possible to use well-known mathematical methods, the choice of which belongs to the project manager. Further research on the earned value method opens up a large space for using mathematical methods for solving specific engineering problems of operational project monitoring.

References

1. Abba, W.: Earned value management-reconciling government and commercial practices. *Program Manager* **26**, 58–63 (1997)
2. Kwak, Y., Anbari, F.: History, practices, and future of earned value management in government: perspectives from NASA. *Proj. Manag. J.* **43**(1), 77–90 (2012)
3. Vandevoorde, S., Vanhoucke, M.: A comparison of different project duration forecasting methods using earned value metrics science direct. *Int. J. Project Manage* **24**, 289–302 (2006)
4. Kerzner, H.: *Project Management: a system approach to planning, scheduling and controlling*. Van Nostrand Reinhold, New York (1984)
5. PMI: Project Management Body of Knowledge (2013). <http://www.pmi.org>
6. Archibald, R.D.: *Managing High-Technology Programs and Projects*. Wiley, Chichester (2003)
7. Bagherpour, M., Noori, S.: Cost management system within a production environment: a performance-based approach. *Proc. Inst. Mech. Eng. Part B: J. Eng. Manuf.* **226**, 145–153 (2011). <https://doi.org/10.1177/0954405411404303>
8. Pajares, J., Lopez-Parades, A.: An extension of the EVM analysis for project monitoring: the Cost Control Index and the Schedule Control Index Science Direct. *Int. J. Project Manag.* (2010)
9. Kuehn, U.: EVM. 05 earned value analysis—why am I forced to do it?. *AACE Int. Trans.* (2007)
10. Gasparotti, C.: Application of the earned value method in monitoring of the project cost. *Rev. Manag. Econ. Eng.* **13**(3)(53), 574–588 (2014)
11. Liu, G., Jiang, H.: Performance monitoring of project earned value considering scope and quality. *KSCE J. Civ. Eng.* **24**(1), 10–18 (2019). <https://doi.org/10.1007/s12205-020-1054-6>



A Tube-Based MPC Structure for Fractional-Order Systems

Stefan Domek^(✉) 

West Pomeranian University of Technology in Szczecin,
ul. Sikorskiego 37, 70-313 Szczecin, Poland
stefan.domek@zut.edu.pl

Abstract. In the paper a new structure to improve the robustness of Fractional-Order Model Predictive Control (FOMPC) are proposed. The method based on the modified Model Following Control (MFC) idea, with the Internal Model Control (IMC) concept, introduced by Skoczowski and Domek in [14]. This leads to a novel, tube-based fractional-order robust predictive control structure, named TFOMPC, which offer an additional degree of freedom in tuning a control loop for higher efficiency. It seems that the proposed TFOMPC approach has potentially great advantages, is simple to implement and easy to tune. Thanks to this, it can be used in many control systems of difficult, inaccurately identified objects, not only of a fractional-order.

Keywords: fractional calculus · fractional-order dynamic models · Model Predictive Control · Model Following Control · tube-based MPC

1 Introduction

The idea of MPC is considered to be, after many years of operating experience in industry, as one of the most universal and effective control methods [8, 17, 18]. It can handle in natural way multivariable systems and moreover, it can take into account explicitly constraints imposed on input and output signals. Along with the development of the theory of fractional differential-integral calculus [5, 12] and the resulting expansion of its practical applications it was also proposed to use the fractional calculus for synthesis of new control methods.

The description of a controlled plant properties by means of fractional-order models can be used indirectly for tuning or directly for design of linear control algorithms. In the second case, fractional calculus has been applied to control theory, which, in its turn, should contribute to the development of new control algorithms significantly different from the well-known integer-order algorithms, and thus, by implication, provide potentially new opportunities for control performance and robustness [10]. Allowing integration/differentiation of arbitrary orders in classic control algorithms results in increasing the number of degrees of freedom in control-parameter tuning, and thus creates new potentialities for control performance and robustness. Excellent examples here are the CRONE algorithm [11] and fractional-order digital $PI^\lambda D^\mu$ algorithm [12], already regarded as

standard, but also the fractional-order iterative learning control, linear-quadratic control, dead-beat control and sliding mode control, which have been proposed in subsequent years, as well as fractional-order model predictive control [3, 13].

In FOMPC the fractional calculus may be applied directly both to defining the cost function and to plant model selection. The use of fractional calculus in the definition of the cost function, although it brings a new degree of freedom in the design of the MPC algorithm, is in many cases criticized by automation specialists, both theoreticians and practitioners, due to the lack of a clear interpretation of the physical criterion of control optimality in a finite prediction horizon. This can be treated more as a potential possibility to modify the form of MPC algorithms than as a change justified in practice and bringing significant differences from the application side [13]. In turn, as is commonly known, the accuracy of the used process model determines directly the MPC quality and therefore cannot be underestimated. Especially when dealing with difficult, inaccurately identified nonlinear or nonstationary plants. The problem is even more noticeable if the controlled plants have behavioral properties that exhibit features inherent in fractional-order models. Therefore, because a model-plant mismatch is inevitable in practice, it leads to deterioration of the MPC control quality, especially during transient processes caused by changes in the operating point or large disturbances. In MPC methods, the natural and most powerful method of reducing negative influences model-plant mismatches is the moving horizon principle. However, its effectiveness is limited in the difficult cases mentioned above. Therefore, providing appropriate system robustness for all objects from the assumed set of uncertainty becomes essential [4].

One of the effective method to improve the robustness of the controller can be an incorporating into a control structure the MFC system elements, which was described closer for the first time by Skoczowski and Domek in [14]. Here the basic control task is performed by the main controller matched in a most optimal way to the nominal process model. On the other hand, the task set for the auxiliary controller is to support the main controller by generating an ancillary signal that depends on the difference between the outputs produced by the adopted model and the actual process and on unknown disturbances. By this means the effect produced by the process-model mismatch (caused, e.g. by different structures) and by possible process perturbations or nonlinearities can be neutralized. The system robustness to model inadequacy, as well as the control performance is thereby increased, and the effect produced by nonmeasurable disturbances is reduced. The principal virtues displayed by the MFC structure are its universality, due to the feasibility of employing arbitrary control algorithms and possibility to design controllers by familiar methods [2, 15].

The paper proposes a new, robust, tube-based fractional-order predictive control structure (TFOMPC). The structure is an extended variant of the MFC concept, with FOMPC as the main controller and a simple proportional ancillary controller. This provides an additional degree of freedom to tune the control loop for greater effectiveness. The proposed control structure is similar to the so-called tube-based MPC for integer-order plants [1, 6, 7, 9, 16].

The paper is organized as follows. After the short introduction in Sect. 1, Sect. 2 reminds the fractional-order differential calculus and fractional-order dynamic models. The Sect. 3 describes the MPC strategy for fractional-order systems and its sensitivity to model-plant mismatch. The proposed robust TFOMPC structure is described in Sect. 4. Finally, Sect. 5 shows some simulation experiments with the TFOMPC structure and various model-plant mismatches, and Sect. 6 concludes the paper.

2 Basics of Fractional-Order Calculus

Differential calculus of non-integer order is a generalization of the classic differential calculus. For a derivative of non-integer order $\alpha \in R$ of a real-valued function $f(t)$, $t \in R$ on the interval $[t_0, t]$, denoted by the operator ${}_{t_0}D_t^\alpha f(t)$, there exist many definitions proposed by various researchers, for example, by Riemann and Liouville ${}^{RL}D_t^\alpha f(t)$, Caputo ${}^C D_t^\alpha f(t)$, Weyl, Fourier, Cauchy, Abel and other. The definitions differ in their properties and/or area of applicability [5, 12]. In practical applications, especially in digital control systems where discrete values of the function $f(kT_s)$, $k = 0, 1, \dots$ taken with the sampling interval T_s are used in a natural way for computations, the most commonly encountered is the definition introduced by Grünwald and Letnikov - a derivative of fractional-order $\alpha \in R$ of the function $f(t)$, $t \in R$, on the interval $[t_0, t]$, is given by

$${}_{t_0}^{GL}D_t^\alpha f(t) = \lim_{T_s \rightarrow 0} T_s^{-\alpha} \sum_{j=0}^{\lfloor \frac{t-t_0}{T_s} \rfloor} c_j^\alpha f(t - jT_s), \quad (1)$$

where the symbol $\lfloor \cdot \rfloor$ denotes the integer part and

$$c_j^\alpha = (-1)^j \binom{\alpha}{j}, \quad \binom{\alpha}{j} = \begin{cases} 1 & \text{for } j = 0 \\ \prod_{i=0}^{j-1} \frac{\alpha-1}{i+1} & \text{for } j = 1, 2, 3, \dots \end{cases} \quad (2)$$

Analyzing the function $f(t)$, $t \in R$ at discrete time-instants kT_s , $k = 0, 1, 2, \dots$ it is also possible to introduce a difference calculus of fractional-order $\alpha \in R$. The equivalent of the derivative (5) in differential calculus is in this case a discrete difference of fractional-order $\alpha \in R$ of the function $f(t)$, $t \in R$, on the interval $[0, kT_s]$, $k \in Z$, is given by

$${}_0\Delta_{kT_s}^\alpha f(kT_s) = T_s^{-\alpha} \sum_{j=0}^k c_j^\alpha f(kT_s - jT_s). \quad (3)$$

Note that the Grünwald-Letnikov derivative of the order $\alpha \in R$ of the function $f(t)$ for discrete-time instants $t_0 = 0$ and $t = kT_s$, $k \in Z$, is given by

$${}_{t_0}^{GL}D_t^\alpha f(t) \Big|_{t=kT_s} = \lim_{T_s \rightarrow 0} ({}_0\Delta_{kT_s}^\alpha f(kT_s)), \quad (4)$$

For discrete-time functions $f(k), k \in Z$, the discrete difference of fractional-order $\alpha \in R$ of the discrete-time function $f(k)$, on the interval $[0, k]$, denoted by the operator ${}_0\Delta_k^\alpha f(k)$, is given by

$${}_0\Delta_k^\alpha f(k) = \sum_{j=0}^k c_j^\alpha f(k-j) \tag{5}$$

and in the simplified form as $\Delta^\alpha f(k) = {}_0\Delta_k^\alpha f(k)$.

Let us consider a continuous-time, nonlinear, fractional-order MIMO system with different, in general, fractional-orders $\Upsilon : \{\alpha_1, \alpha_2, \dots, \alpha_n\}$ for individual state variables x_1, x_2, \dots, x_n

$${}^G L_0 D_t^\Upsilon x(t) = f(x(t), u(t)) + v(t) \tag{6}$$

$$y(t) = g(x(t), u(t)) + d(t), \quad t \in R \tag{7}$$

where $u(t) \in R^m$, $y(t) \in R^p$ denote the input and output vectors and $v(t) \in R^n$, $d(t) \in R^p$ are the state and output disturbance vectors, respectively, while

$${}^G L_0 D_t^\Upsilon x(t) = [{}^G L_0 D_t^{\alpha_1} x_1(t) \quad {}^G L_0 D_t^{\alpha_2} x_2(t) \quad \dots \quad {}^G L_0 D_t^{\alpha_n} x_n(t)] \tag{8}$$

is the generalized fractional derivative of the state vector $x(t) \in R^n$.

Taking into account (4), (5) and (8) we can formulate a generalized linear, discrete fractional-order model in the state space of the MIMO plant (6), (7), (8):

$$x(k+1) = Ax(k) + Bu(k) - \sum_{j=1}^{k+1} \Upsilon_j x(k+1-j) + v(k) \tag{9}$$

$$y(k) = Cx(k) + d(k), \quad k \in Z \tag{10}$$

where the state matrix $A \in R^{n \times n}$, input matrix $B \in R^{n \times m}$ and output matrix $C \in R^{p \times n}$ are the corresponding derivatives of functions $f(x(t), u(t))$, $g(x(t), u(t))$ at the operating point of a nonlinear process and the matrices of coefficients (2) for individual fractional orders

$$\Upsilon_j = \text{diag} [c_j^{\alpha_1} \quad c_j^{\alpha_2} \quad \dots \quad c_j^{\alpha_n}] \in R^{n \times n}. \tag{11}$$

In practical applications of model (9)–(11) it is not possible to take for numerical calculations the state vector $x(k+1-j)$ samples, the number of which grows rapidly with increasing discrete time k . One of the many methods to cope with the problem is to adopt a finite-length memory in which the instantaneous values of the state vector are stored. Let's note that the generalized fractional discrete difference of the state vector $x(k)$ with different orders for individual state variables (12) can be written as

$$\Delta^\Upsilon x(t) = \sum_{j=0}^L \Upsilon_j x(k-j) + [\mathcal{R}_L^{\alpha_1} \quad \mathcal{R}_L^{\alpha_2} \quad \dots \quad \mathcal{R}_L^{\alpha_n}]^T \tag{12}$$

where L is the adopted memory length, and

$$\mathcal{R}_L^{\alpha_i} = \sum_{j=L+1}^k c_j^{\alpha_i} f(k-j) x_i(k-j) \quad \text{for } i = 1, 2, \dots, n \quad (13)$$

is the so-called tail which can be made arbitrarily small with an appropriate choice of L , although, in practice large values of L should be avoided to limit the complexity of numerical calculations [4].

Taking above into account, according to the approach taken from the theory of FIR digital filters, the so-called truncated generalized fractional Grünwald-Letnikov difference, very close to the true fractional difference value (12), can be obtain:

$$\Delta_L^{\mathcal{X}} x(k) = [\Delta_L^{\alpha_1} x_1(k) \quad \dots \quad \Delta_L^{\alpha_n} x_n(k)]^T = \sum_{j=0}^L \Upsilon_j x(k-j) \approx \Delta^{\mathcal{X}} x(k) \quad (14)$$

with the assumption that the upper limit of summation must be reduced to the value of (k) until enough samples are accumulated. Consequently, the discrete, fractional-order model (9), takes the form with finite-length memory:

$$x(k+1) = Ax(k) + Bu(k) - \sum_{j=1}^L \Upsilon_j x(k+1-j) + v(k). \quad (15)$$

Note that the solution of the state equation (15) is known and additionally, a discrete model in a form of a transfer functions matrix $M(z^{-1})$ can be obtain [5, 12].

3 Fractional-Order Model Predictive Control

In fractional-order MPC for a fractional-order system (6)–(8), the optimal sequence of future increments of control actions $\Delta u(k+j|k)$ are to be found at each instant $k \in Z$ within the control horizon from $j = 0$ to $j = N_u - 1$, to minimize the performance index along horizons of future time instants

$$V(k) = [Y^r(k) - Y^p(k)]^T M [Y^r(k) - Y^p(k)] + [\Delta U(k)]^T \Lambda [\Delta U(k)], \quad (16)$$

s.t.

$$u_{min} \leq u(k+j|k) \leq u^{max}, \quad j = 0, 1, \dots, N_u - 1, \quad (17)$$

$$y_{min} \leq y^p(k+j|k) \leq y^{max}, \quad j = N_1, N_1 + 1, \dots, N_2, \quad (18)$$

where M and Λ are symmetric, positive semidefinite and positive definite weighting matrices, most often ($M = I_{(N_2-N_1+1) \cdot p}$, $\Lambda = \lambda I_{N_u \cdot m}$), and the vectors $Y^r(k)$, $Y^p(k)$ group future values of reference input and model output in the prediction horizon $j \in [N_1, N_2]$, respectively.

The predicted response of the model (15), needed to numerically solving of the optimization task (16)–(18) can be determined from the solution of the state equation [5]:

$$y^p(k+j|k) = C \left[\Phi^X(j) x(k) + \sum_{i=0}^{j-1} \sum_{l=0}^i \Phi^X(j-l-1) B \Delta u(k+l) + \sum_{i=2}^L \sum_{l=-1}^{-i+1} (-1)^{i+1} \Phi^X(j-l-i) x(k+l) + v(k+j|k) \right] + d(k+j|k) \quad (19)$$

where $v(k+j|k)$ and $d(k+j|k)$ may be assessed just as the differences between the measured (estimated) state and output and the state and output predicted for the current sampling instant at the sampling instant $k-1$ respectively, what are most popular but also quite simple and known as the DMC-type models of disturbances [2, 17].

FOMPC, as written, can handle in a natural way multivariable systems, and moreover, it can take into account explicitly various signals constraints and various kinds of disturbances. However, its effectiveness depends on the accuracy of the process model that is utilized directly to compute the manipulated variable [4, 16]. Especially when dealing with difficult, inaccurately identified, nonlinear or nonstationary plants. The problem is even more noticeable if the controlled plants have behavioral properties that exhibit features inherent in fractional-order systems. In FOMPC, as well as, in integer-order MPC, the natural and most powerful method of reducing negative influences model-plant mismatches is the moving horizon principle. That means only the first value of the computed control sequence is fed as the input into the real plant, and the whole procedure is repeated at the following discrete time instants, whereas the reference trajectory $Y^r(k)$ starts each time from the current output value of the plant $y(k)$. Thanks to this, prediction errors resulting from insufficient quality of the object model and models of unmeasurable disturbances as well as possible numerical errors can be compensated on an ongoing basis. However, effectiveness of moving horizon method is limited in the difficult cases mentioned above [2].

Effects of the model-plant mismatch on the effectiveness of FOMPC algorithms was considered in [4]. Results of experiments for various types of mismatches were show, e.g. uncertainties of matrices A, B , fractional orders \mathcal{Y} and memory lengths L in model (15), as well as uncertainty of a matrix C in (10). The results confirmed the noticeable sensitivity of the FOMPC system to the model-plant mismatch.

4 Proposed a Tube-Based FOMFC Structure

An effective method to improve the robustness of control systems is the use of the so-called Model Following Control (MFC) approach. Generally, in MFC systems the basic control task is performed by the main controller matched in a most optimal way to the nominal process model and the task set for additional

elements is to support the main controller by generating an ancillary signal. It depends on the difference between the outputs produced by the adopted model and the actual process and on unknown disturbances. By this means, the effect produced by the process-model mismatch and by possible process perturbations can be neutralized, and the control performance is thereby increased. Figure 1 shows the considered robust control structure [2, 14].

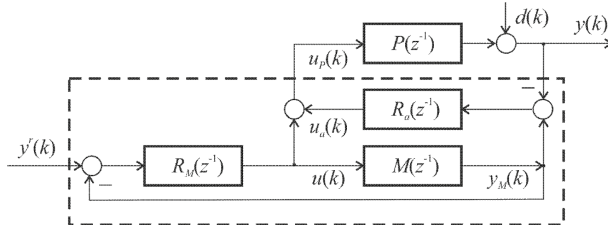


Fig. 1. Proposed robust FOMPC structure

Properties exhibited by the MFC structure are described most often by treating perturbations as a synthetic description of changes of all kinds experienced by the process in reference to the nominal model, thus indirectly as a description of the mismatch between the adopted truncated nominal model (15) and the actual fractional-order nonlinear and/or nonstationary process (6)–(8). It is assumed that the system components are described by discrete transfer function matrices of appropriate dimensions: main controller $R_M(z^{-1})$, ancillary controller $R_a(z^{-1})$, nominal model $M(z^{-1})$ and process $P(z^{-1})$. Next, it is assumed that the process is subjected to multiplicative perturbations $\Delta(z^{-1})$ in relation to the nominal process represented by its model (15) in the discrete transfer function form.

In [2], the robustness of the MFC structure was compared to the robustness of a single-loop feedback control structure. Among others, functions of disturbance sensitivity and input sensitivity in the frequency domain was checked. It has been shown that the disturbance suppression in the MFC system compared to the classical system is easier. It was also shown which maximum boundary perturbations are acceptable without the risk of loss of stability in the MFC structure. In conclusion of [14], some important remarks about methods for tuning component controllers operating in the MFC system have been formulated:

- all known methods for tuning the main controller operating in classic single-loop feedback systems are applicable. Note that the main controller governs a model with known and constant parameters, hence, its tuning is relatively simple. In addition, the accuracy of the process model identification is not a crucial matter here owing to the inherent robustness of the MFC structure;
- the auxiliary controller gain should be as high as possible in order to extend the range of allowable process perturbations, with due regard for stability conditions that impose a bound from above. It is worth noting that the reference

signal for the auxiliary controller, i.e. the model output $y_M(k)$, is smoothed by the main controller. Additionally, the auxiliary controller is supported by the main controller in that it works out the initial value of the manipulated variable $u(k)$. For all these reasons, the auxiliary controller operates under much more favorable conditions and thus from a practical point of view should be simple. Moreover, it may be tuned to a smaller stability margin than controllers in the classic single-loop structure.

Considering that the main advantages of the MFC structure are its robustness, universality due to the feasibility of employing arbitrary control algorithms, and the possibility of designing controllers using known methods, in this paper we propose a new robust fractional-order MPC. It is based on the robust MFC structure described above and is designed as follows:

- the main controller R_M is formulated as FOMPC (16) subject to constraints (17), (18);
- as the internal model for R_M the fractional-order truncated model (10), (11), (15) without disturbances $d(k)$ and $v(k)$ are chosen;
- the auxiliary controller R_a is proportional with the gain matrix K :

$$u_a(k|k) = K(y_M(k|k) - y(k)); \quad (20)$$

- the constraint (17) is modified and take the form:

$$u_{min} - u_a(k|k) \leq u(k+j|k) \leq u^{max} - u_a(k|k), \quad j = 0, 1, \dots, N_u - 1. \quad (21)$$

As a result, the new robust TFOMPC is obtained which is similar to the concept of so-called tube-based MPC known from literature [1, 6, 7], [9, 16].

5 Numerical Examples

Consider the fractional-order plant (9)–(11) with

$$A = \begin{bmatrix} 2.7756 & -1.2876 & 0.7985 \\ 2.0000 & 0 & 0 \\ 0 & 0.5000 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0.0313 \\ 0 \\ 0 \end{bmatrix}, \quad C = [1 \ 0 \ 0], \quad \alpha = 0.8.$$

This plant will be perturbed in fractional orders = 0.7, 0.8, 0.9. For the set-point tracking, first the FOMPC controller (16) with $N_1 = 1$, $N_2 = 10$, $N_u = 2$, $M = I$, $\Lambda = 5I$, $T_p = 1$ s, has been used. Figure 2 shows the step responses in the control system for the truncated fractional-order model (15), with memory-lengths ($L_M = 100$). Next, the proposed TFOMPC with the same internal nominal model (15) ($\alpha = 0.8$, $L_M = 100$) was used. Figure 3 shows the step responses in the control system for $K = 0.1$. It can be noticed that the controller is much more robust to perturbations of the plant.

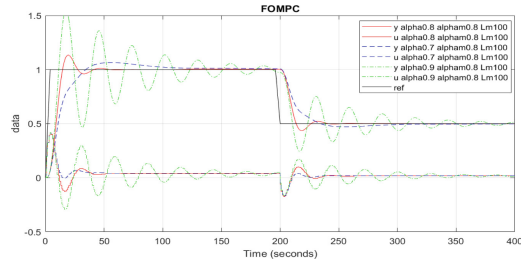


Fig. 2. Step responses in the FOMPC control system

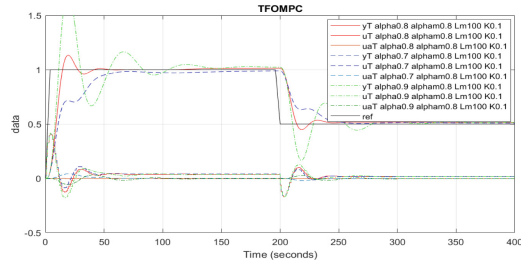


Fig. 3. Step responses in the proposed TFOMPC control system

6 Conclusion

The paper presents a new robust fractional-order MPC structure based on the well-known Model Following Control scheme. The obtained control structure is similar to the concept of the so-called tube-based MPC, which is often considered in recent literature. The proposed TFOMPC structure is easy to synthesis and implementation, and can be used in many difficult and perturbed systems of fractional-order. Its effectiveness was demonstrated by some simulation examples. Future work may include studies of modified TFOMPC structures, such as the MFC case with the model state zamiast output vector and process state estimation in the cost function (16) and the auxiliary controller (20). The benefits of the practical implementation of the proposed structure should also be assessed more closely.

References

1. Brunner, F.D., Müller, M.A., Allgöwer, F.: Enhancing output feedback MPC for linear discrete-time systems with set-valued moving horizon estimation. In: Proceedings of IEEE Conference on Decision and Control (2016)
2. Domek, S.: Robust Model Predictive Control for Nonlinear Processes, vol. 593. Technical University of Szczecin Academic Press, Szczecin (2006). (in Polish)
3. Domek, S.: Multiple use of the fractional-order differential calculus in the model predictive control. In: 19th International Conference on Methods and Models in Automation and Robotics (MMAR), pp. 359–362 (2014)

4. Domek, S.: Model-plant mismatch in fractional order model predictive control. In: Domek, S., Dworak, P. (eds.) *Theoretical Developments and Applications of Non-Integer Order Systems*. LNEE, vol. 357, pp. 281–291. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-23039-9_24
5. Kaczorek, T.: *Selected Problems of Fractional Systems Theory*. LNCIS, Springer, Berlin (2011). <https://doi.org/10.1007/978-3-642-20502-6>
6. Kögel, M., Findeisen, R.: Robust output feedback MPC for uncertain linear systems with reduced conservatism. *IFAC-Papers Online* **50**, 10685–10690 (2017)
7. Lorenzetti, J., Pavone, M.: A simple and efficient tube-based robust output feedback model predictive control scheme. In: *European Control Conference (ECC)*, pp. 1775–1782 (2020)
8. Ławryńczuk, M.: *Computationally Efficient Model Predictive Control Algorithms*. SSDC, vol. 3. Springer, Cham (2014). <https://doi.org/10.1007/978-3-319-04229-9>
9. Mayne, D.Q., Raković, S.V., Findeisen, R., Allgöwer, F.: Robust output feedback model predictive control of constrained linear systems. *Automatica* **42**, 1217–1222 (2006)
10. Monje, C.A., Chen, Y.Q., Vinagre, B.M., Xue, D., Feliu, V.: *Fractional Order Systems and Controls*. Springer, London (2010). <https://doi.org/10.1007/978-1-84996-335-0>
11. Oustaloup, A.: *La Derivation Non Entiere: Theorie. Synthese et Applications*, Hermes, Paris (1995)
12. Podlubny, I.: *Fractional Differential Equations*. Academic Press, San Diego (1999)
13. Romero, M., De Madrid, Á.P., Mañoso, C., Vinagre, B.M.: Fractional-order generalized predictive control: formulation and some properties. In: *11th International Conference on Control, Automation, Robotics and Vision*, pp. 1495–1500 (2010)
14. Skoczowski, S., Domek, S.: Robustness of a model following control system. In: *Proceedings of International Conference on Mathematical Theory of Networks and Systems (MTNS)*, CD (2000)
15. Skoczowski, S., Domek, S., Pietruszewicz, K., Broel-Plater, B.: A method for improving the robustness of PID control. *IEEE Trans. Ind. Electron.* **52**, 1669–1676 (2005)
16. Sotasakis, P., Sarimveis, H.: Stabilising model predictive control for discrete-time fractional-order systems. *Automatica* **75**, 24–31 (2017)
17. Tatjewski, P.: *Advanced Control of Industrial Processes*. Springer, London (2007). <https://doi.org/10.1007/978-1-84628-635-3>
18. Tatjewski, P.: Disturbance modeling and state estimation for offset-free predictive control with state-spaced process models. *Int. J. Appl. Math. Comput. Sci.* **24**, 313–323 (2014)



Creep Testing Machine Identification for Power System Load Optimization

Michał Szulc^{1,2}(✉), Jerzy Kasprzyk², and Jacek Loska^{1,2}

¹ Łukasiewicz Research Network – Upper Silesian Institute of Technology, Gliwice, Poland
michal.szulc@git.lukasiewicz.gov.pl

² Silesian University of Technology, Gliwice, Poland

Abstract. This paper presents the results of an identification study of a three-zone furnace of a single-sample creep testing machine. The steps leading to obtaining a model are described, including: preparation of the test stand, determination of parameters of the identification experiment, identification of the dominant disturbances, selection of excitation signal, selection of model structure, validation of the obtained results, and analysis of possible nonlinearities. The description of the research includes information on the applied software compensator for disturbances caused by changes in the supply voltage, with the help of which they were reduced to a level that made it possible to conduct identification experiments. A model with three inputs and three outputs was identified as a MIMO, MISO and SISO structure. The advantages and disadvantages of each of these solutions are presented. The model was validated using step response experiments. The paper also presents a method for including static nonlinearity in the model.

Keywords: Creep test · System identification · Dynamic model · Disturbance compensation

1 Introduction

The creep test laboratory, located at the Upper Silesian Institute of Technology in Gliwice, is used to test the strength of steel materials for steel producers and users, as well as all institutions related to the iron and steel industry. It includes 87 one-sample and four multi-sample creep testing machines. The temperature in the furnaces of the creep testing machines is stabilized using electric heaters controlled by on-off controllers. In the case of a mains power failure it is powered by an emergency generator. In total, 389 heaters are switched on in the laboratory, and unsynchronised switching generates large power peaks that prevent proper operation of the emergency generator. Among several solutions to the problem, it was decided to develop a master control algorithm to minimise the power peaks caused by unsynchronised switching. Since the laboratory work cannot be interrupted for the duration of research on the control algorithm, work on the search for an appropriate algorithm will be carried out in the simulation mode. Thus, there is a need to create a mathematical model describing the behaviour of a single test machine. Such a model should be relatively simple, because it will be necessary to

simulate 91 creep testing machines. Due to the difficulty of creating such a model by analysing the phenomena occurring in the process, it was assumed that it can be obtained by identification.

The paper is organised as follows: (1) description of the test stand, (2) presentation of the identification experiment and results of identification, (3) validation of the obtained models, and final conclusions.

2 Description of the Test Stand

The test object was a three-zone resistance furnace in the shape of a cylinder with a diameter of 228 mm and a length of 260 mm. A 54 mm diameter ceramic tube was placed in the axis of the cylinder, in which the test sample was located. Three heating coils were equally spaced along the length of the outer wall of the ceramic tube. The nominal power of each heater was 1300 W. The set power of the heaters was expressed as a percentage of the nominal power, which was converted to the time the heater was switched off, with the constant time it was switched on.

The temperature was measured using three S-type thermocouples attached to a standard axial sample with a measuring length of 50 mm. Two of the thermocouples were placed at the beginning and the end of the measuring length of the sample, while the third was placed in the middle of the sample. The temperature was measured with a 16-bit resolution. The cold junction compensation of the thermocouples was achieved using a Pt100 sensor. Data acquisition was carried out with a sampling interval of 1 s. The test stand with selected components is shown in Fig. 1.

The identification experiment was carried out using an available creep testing machine, without any hardware modifications. The specimen was placed in the coupler shanks and then, after being placed in the furnace, connected to the lever loading system. The preparation of the test bench for the identification experiment was carried out according to the creep test preparation procedure. The thermocouples and measuring channels were calibrated at the test temperature, and the resulting systematic errors were used as a correction to the measured value. The measurement uncertainty was 0.9 °C. Immediately before the experiment started, the sample was loaded. The furnace

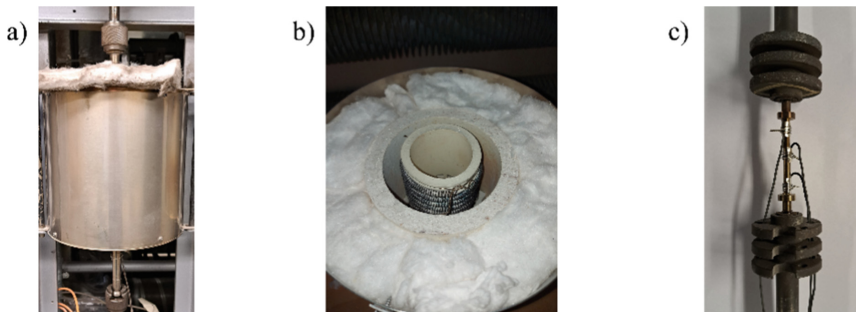


Fig. 1. The test stand incl. a) creep testing machine, b) int. construction of furnace, c) sample with thermocouples attached

was insulated with ceramic plugs and insulating glass wool. An insulating covering was placed on top of the furnace. The creep test room was equipped with a ventilation and air-conditioning system to maintain a temperature of 23 °C with an accuracy of ± 3 °C around each test bench.

3 Identification Experiment

The identification experiment was preceded by an analysis of the influence of disturbances on the process and accuracy of the model, their compensation, initial evaluation of the dynamics in order to select the sampling period and the appropriate excitation signal [1, 2].

Model identification requires the power of the useful output signal to be significantly greater than the power of the disturbance. On the other hand, due to the potential nonlinearities of the object, a low amplitude output signal is required. Based on the initial analysis of the process, the maximum value of the object's excitation response amplitude was set at 10% of the set point. The object operates in a temperature range of 400 to 800 °C. It was assumed that the test would be conducted at a temperature of approximately 600 °C. The response amplitude for this set point was not to exceed 60 °C [2].

The effect of disturbances was measured at constant heater power values. The amplitude of the disturbances was approximately 20 °C. This value prevented correct identification of the model, so it was decided to identify the disturbance source in order to compensate it. The disturbances affecting the test object included changes in environmental temperature, changes in the supply voltage of the furnace heaters, and quantisation noise. The temperature changes in the machine environment were measured using a Pt100 sensor. The results are shown in Fig. 2. The peak-to-peak value of the environmental temperature was about 2.5 °C.

Thus, it was to be expected that its effect on the temperature of the sample would be significantly lesser than the value obtained as the disturbance caused by changes in ambient temperature accounted for only 10% of the total disturbance affecting the test object. The conclusion is that another dominant disturbance affected the object. The influence of quantisation noise was eliminated by low-pass filtering of the output signal. The effect of changes in the supply voltage of the furnace heaters on changes in the temperature of the object was investigated using a power meter. The measurement was performed with the shortest possible sampling period equal to 1 s. Considering that fast changes in the mains voltage should be completely filtered by the test object, the signal obtained was low-pass filtered [2]. The effect of long-term mains voltage fluctuations of the output signal changes is shown in Fig. 3. The observed correlation between the presented signals allows to conclude that the greatest influence on the output signal came from the changes in the mains voltage.

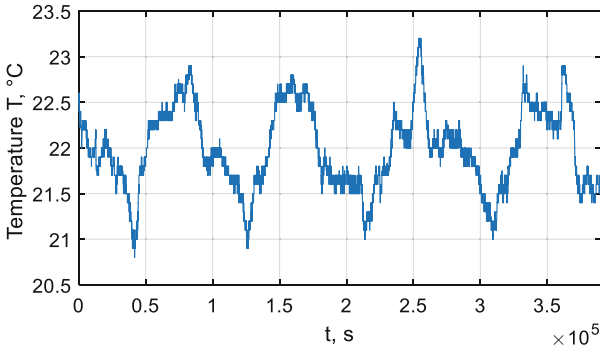


Fig. 2. Environmental temperature

The compensation of the disturbance from the supply voltage was obtained in a programmable way by correcting the time between heater starts. The formula for the turn-off time correction Δt_{offp} was obtained by performing a simple transformation of the power vs. supply voltage and resistance relationship, and using the formula that determined the turn-off time of the heater at a given percentage of the heater’s power $P_{\%}$ at a given fixed turn-on time t_{on} :

$$\Delta t_{offp} = \frac{t_{on}}{P_{\%}} \left(1 - \frac{U_{ref}^2}{U^2} \right), \tag{1}$$

where U and U_{ref} denote the measured and reference voltage, respectively.

The interference amplitude before, and after compensation is shown in Fig. 3. The maximum disturbance amplitude decreased from about 20 °C to about 2.5 °C, which represents about 4% of the disturbance contribution to the output signal which should not have a significant impact on the identification experiment.

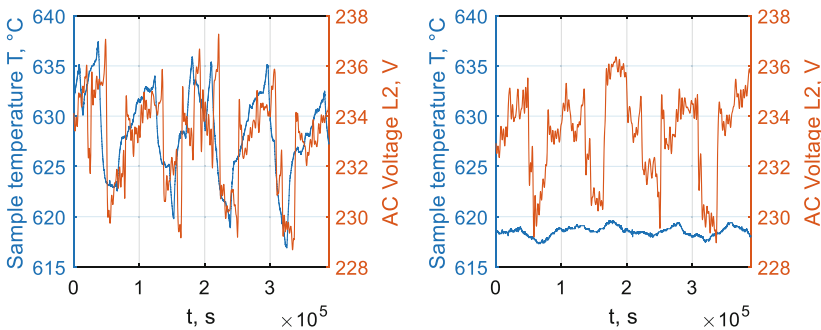


Fig. 3. Effect of changes in mains voltage on sample temperature before, and after compensation

A preliminary assessment of the dynamics of the object was made by estimating the dominant time constant of the step response. This was approximately 7200 s. In addition, a test of the object’s response to a rectangular signal excitation was carried

out by varying its period. As a result of the test, the limiting frequency of the signal carried by the object was determined. Based on the accepted principle that the sampling period should be equal to 10% of the dominant time constant [2], a sampling period of 720 s should have been used. However, due to the range of transferred frequencies, the sampling period of 360 s was selected.

In multi-input systems, the input signals are selected to ensure that there is no correlation between them [3]. In this study, a pseudorandom binary signal (PRBS), having properties similar to white noise, was used for each input [2]. In order to ensure an output signal amplitude of approximately ± 60 °C, an input signal amplitude value of 5% was estimated. The identification experiment was carried out using a tenth order excitation signal of 1023 samples with a sampling period of 360 s.

Due to the lack of an antialiasing filter at the measurement input, it was decided to take raw data from the system with a measurement card period of 1 s. Then, the data was digitally low-pass filtered and the resulting signal was decimated to a sampling period of 360 s. Before model identification, the DC component, whose presence in the signal can lead to erroneous results, was also removed [2].

4 Results of Identification

The system to be identified has 3 inputs, *i.e.* power supply for 3 heaters (upper, middle and lower), and 3 outputs, *i.e.* the temperature of the sample measured at 3 points. This system can be identified as a single MIMO model, as 3 MISO models for 3 outputs, or as 9 SISO models separately for each control path. Different model structures were tested, and the best results were obtained using the stationary ARMAX model [1]:

$$A(z^{-1})y(i) = B(z^{-1})u(i) + C(z^{-1})e(i), \quad (2)$$

where, depending on the dimensionality of the model, A , B and C denote polynomials of the lag operator z^{-1} , or matrices of polynomials with appropriate dimensions, y , u and e are outputs, inputs and white noise at time instant i , respectively.

The parameters of the SISO models are scalars, which facilitates identification. But the disadvantage of this solution is the need to perform nine experiments, and then assembling nine SISO models into one MIMO model [4]. The MIMO structure is most suitable for identifying multidimensional models. The model is obtained by single identification procedure, but the selection of parameters, due to the matrix representation, is much more difficult. A compromise between the SISO and MIMO models is the MISO model. For the object under study, it requires the identification of only three component models, where their parameters are scalars and three-element vectors, respectively [5]. In this study, all three approaches were tested.

The optimal MIMO model was obtained by initially searching for a stable model, and then the structure of the model minimizing the identification error was sought for [6]. This is a very tedious procedure, because in this model as many as 21 polynomials have to be selected, and there are no prerequisites for the structure determination. The model finally chosen is presented in Table 1. A comparison between the response of the final MIMO model and object to the excitation used in the experiment is shown in

Table 1. The resulting MIMO model

Output	Model
y_1	$(1 - 2.89z^{-1} + 1.34z^{-2} - 0.28z^{-3}) y_1(i) =$ $-(2.51z^{-1} - 0.92z^{-2}) y_2(i) - (-1.07z^{-1} + 0.37z^{-2}) y_3(i) +$ $0.54z^{-1} u_1(i) + (1.43z^{-1} + 0.32z^{-2}) u_2(i) + 0.34z^{-1} u_3(i) + (1 - 0.70z^{-1}) e(i)$
y_2	$(1 + 1.68z^{-1} - 0.56z^{-2} - 0.21z^{-3}) y_2(i) =$ $-(-1.79z^{-1} + 0.78z^{-2} - 0.06z^{-3}) y_1(i) - (-1.32z^{-1} + 0.52z^{-2}) y_3(i) +$ $+ 0.37z^{-1} u_1(i) + (1.73z^{-1} + 0.28z^{-2}) u_2(i) + 0.48z^{-1} u_3(i) + (1 - 0.66z^{-1}) e(i)$
y_3	$(1 - 2.25z^{-1} + 1.16z^{-2} - 0.30z^{-3}) y_3(i) =$ $-(-1.12z^{-1} + 0.54z^{-2} - 0.11z^{-3}) y_1(i) - (1.88z^{-1} - 0.91z^{-2} + 0.16z^{-3}) y_2(i) +$ $+ 0.24z^{-1} u_1(i) + 1.45z^{-1} u_2(i) + 0.74z^{-1} u_3(i) + (1 - 1.26z^{-1} + 0.77z^{-2}) e(i)$

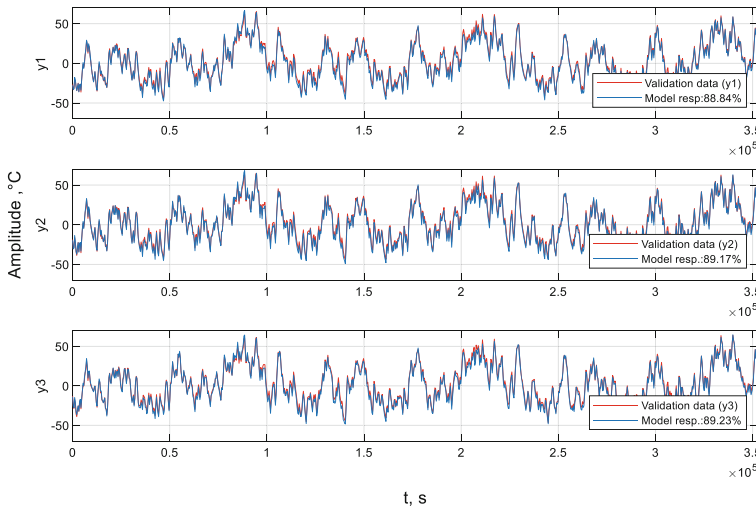


Fig. 4. Comparison of MIMO model response with validation data

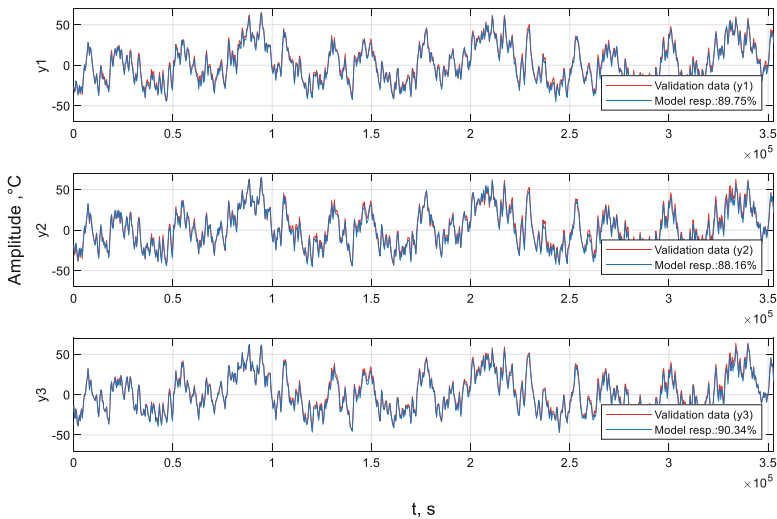
Fig. 4. The obtained model correlates to 93% of the estimated data, giving acceptably good result.

The MISO model was obtained by independently identifying three models, each for a different output. The chosen MISO model is shown in Table 2. A comparison of the response of the object and a model to the excitation used during identification is shown in Fig. 5. Results similar to the MIMO model were obtained, but with fewer steps leading to the model.

SISO models were also identified, but because of the reasons described below, the presentation of the results is omitted.

Table 2. The resulting MISO model

Output	Model
y_1	$(1 - 1.43z^{-1} + 0.56z^{-2} - 0.10z^{-3}) y_1(i) = 0.57z^{-1} u_1(i) + (1.37z^{-1} - 0.33z^{-2} - 0.41z^{-3}) u_2(i) + (0.28z^{-1} + 0.14z^{-2}) u_3(i) + (1 - 0.81z^{-1}) e(i)$
y_2	$(1 - 1.65z^{-1} + 0.94z^{-2} - 0.26z^{-3}) y_2(i) = 0.41z^{-1} u_1(i) + (1.65z^{-1} - 1.04z^{-2}) u_2(i) + 0.44z^{-1} u_3(i) + (1 - 0.82z^{-1}) e(i)$
y_3	$(1 - 1.60z^{-1} + 0.55z^{-2} + 0.14z^{-3} - 0.07z^{-4}) y_3(i) = (0.18z^{-1} + 0.15z^{-2} - 0.12z^{-3}) u_1(i) + (1.38z^{-1} - 0.61z^{-2} - 0.76z^{-3} + 0.30z^{-4}) u_2(i) + (0.64z^{-1} + 0.15z^{-2} - 0.50z^{-3}) u_3(i) + (1 - 0.67z^{-1}) e(i)$

**Fig. 5.** Comparison of MISO model response with validation data

5 Model Validation

The validation of the models was carried out using the step response to a 3% power step. The step responses of the MIMO model compared with measured step responses are shown in Fig. 6. Similar results were obtained for the MISO model.

The dynamics of the model response is similar to that generated by the object. Responses to excitation at input u_3 are characterised by high correlation, in other cases the model gain is lower than the gain revealed by the object step response. It may suggest that the object under consideration is nonlinear. This is rather obvious, since the dynamics in the direction of heating differs from the dynamics in the direction of cooling. Nevertheless, for testing the power distribution control algorithm between individual machines, it should not be of great importance, due to the short switching times. Therefore, it can be assumed that the obtained models meet the accuracy requirements.

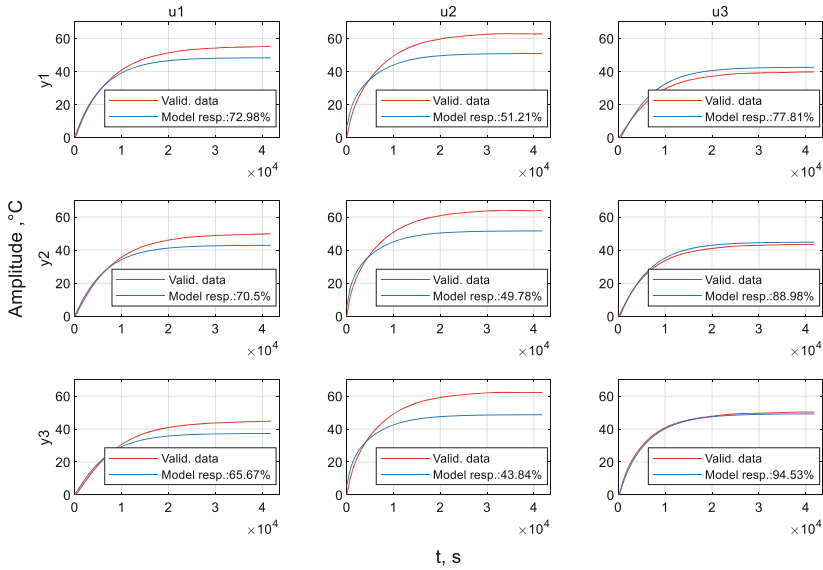


Fig. 6. Comparison of MIMO model response to a 3% step of power with validation data

However, there is also another source of non-linearity. The model was designed for a temperature around 600 °C. Meanwhile, it is known that the higher the temperature, the lower the gain for heating. Hence, an additional experiment was performed consisting in determining the static characteristics of the entire range of temperatures occurring in the machine. The result of approximation the measurement points with a fourth-order polynomial is shown in Fig. 7.

This made it possible to introduce compensation for this non-linearity as in the Wiener model [7]. The use of the compensation function allowed to extend the applicability of the model to the entire range of the machine’s operation, and also improved the model at the bias point, as shown in Fig. 8.

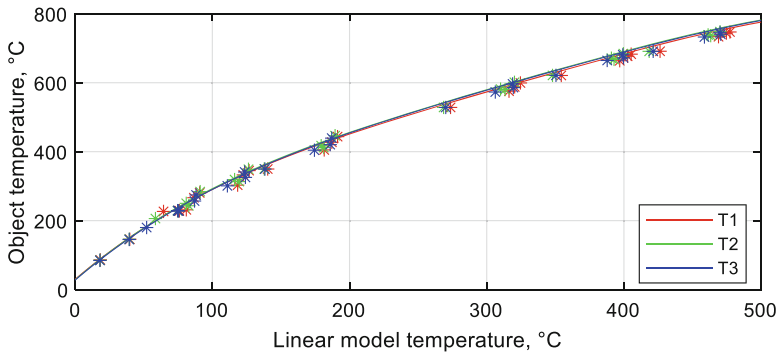


Fig. 7. Nonlinearity compensation function

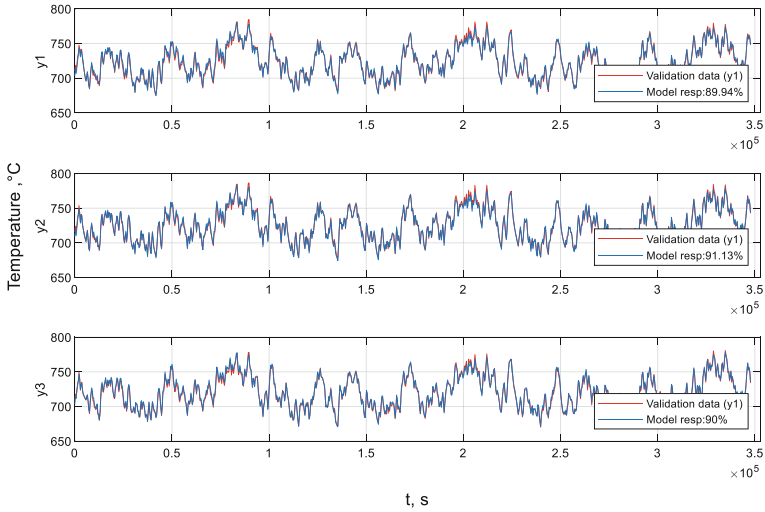


Fig. 8. Comparison of MIMO model responses with validation data after nonlinearity compensation

6 Conclusions

The presented results of the identification of the furnace in a one-sample creep testing machine were used to build a simulator of a creep test laboratory. The models were sufficiently accurate to achieve the aim of searching for a master control algorithm that optimizes the phase load in a multi-load system. The best model was obtained as the MIMO ARMAX model with a static polynomial model added at the output. However, creating this model required more effort than with the MISO or SISO structures. This model also performed better in the case of static non-linearity compensation than the MISO and SISO models, where the introduction of non-linear compensation significantly deteriorated their accuracy at the selected operating point. However, this compensation did not eliminate the errors introduced by other sources of non-linearity, such as those, resulting from differences in heating and cooling time constants, as well as the effect of radiation on sample heating. In addition, the presence of disturbances forced the identification experiment to be carried out with a fairly large amplitude of the output signal, outside the linearity range of the object, which also affected the accuracy of the obtained models.

References

1. Ljung, L., Torkel, G.: Modeling of Dynamic Systems. Prentice Hall Information and System Sciences Series. PTR Prentice Hall, Englewood Cliffs (1994)
2. Ljung, L.: System Identification: Theory for the User. Prentice Hall Information and System Sciences Series, 2nd edn. PTR Prentice Hall, Upper Saddle River (1999)
3. Figwer, J., Niederliński, A., Kasprzyk, J.: A new approach to the identification of linear discrete-time MISO systems. Arch. Control Sci. **2**(3–4), 223–237 (1993)

4. Isermann, R., Münchhof, M.: Identification of dynamic systems: an introduction with applications (2011). <https://doi.org/10.1007/978-3-540-78879-9>
5. Shreesha, C., Gudi, R.D.: MISO structure based control-relevant identification of MIMO systems. In: Proceedings of the American Control Conference, pp. 1184–1189 (2001)
6. Kasprzyk, J.: Model structure determination in parametric model identification. *Syst. Sci.* **23**(2), 88–95 (1997)
7. Bai, E.: A blind approach to Hammerstein–Wiener model identification. *Automatica* **38**(6), 967–979 (2002)



Development and Testing of the RFID Gripper Prototype for the Astorino Didactic Robot

Adrian Kampa¹(✉), Krzysztof Foit¹, Agnieszka Sękala¹, Jakub Kulik¹, Krzysztof Łukowicz¹, Miłosz Mróz¹, Julia Nowak¹, Marek Witański¹, Patryk Żebrowski¹, Tomasz Błaszczyk², and Dariusz Rodzik³

¹ Silesian University of Technology, Gliwice, Poland
adrian.kampa@polsl.pl

² Technical University of Denmark, Lyngby, Denmark

³ Military University of Technology, Warszawa, Poland

Abstract. The rapid development of robotics creates the need for systematic training and professional competence of workers, which caused many projects of teaching robots to be developed. Industrial robots are versatile manipulation machines, but they need appropriate tools and workstation equipment to perform their tasks. As part of the student PBL (Project Based Learning) project, a concept was developed and a prototype teaching workstation containing the Astorino educational robot with equipment was made. Elements of the workstation's equipment enable the demonstration of selected capabilities of Industry 4.0, such as the Internet of Things and systems allowing RFID wireless identification of manipulated objects, among others. As part of the PBL project, gripper jaws integrated with an RFID sensor were developed, and identification range tests were conducted for different gripper configurations and manipulated objects with different types of RFID labels. The results show that the tested labels have a small read range of about 20–40 mm, which is related to the limitations of the 13.56 MHz HF RFID technology used, as well as the dimensions of the labels and the materials used. This parameter is of great importance in determining the applicability of RFID labels, so the results of the study will be used for further development work on RFID grippers.

Keywords: automation and robotics · mechanical engineering · Industry 4.0 · Internet of Things · RFID - radio frequency identification

1 Introduction

The increasing robotization and implementation of the Industry 4.0 concept require the training of robotic engineers and robot operators. Robotics and related courses are present in many schools and universities to prepare students and pupils for work with industrial robots and to train them in the modern technologies of Industry 4.0 [9], including the Internet of Things (IoT) and Radio-Frequency Identification (RFID) technology [18, 20]. With the mentioned needs in mind, a student PBL (Project Based Learning)

project related to development work on the Astorino robot prototype [21], provided by Astor, was undertaken at the Silesian University of Technology. As part of the project, a concept was developed and a prototype station was fabricated to enable the development work. Elements of the workstation equipment, such as manipulated objects, magazines, a trajectory plotting board, a pen attachment tool, and a pneumatic gripper adapter were designed and fabricated. One of the tasks completed in the project was the development of a prototype and testing of the RFID gripper for the Astorino teaching robot, which is presented in the following sections.

2 Literature Review

Industry 4.0 is a multifaceted concept that integrates various technologies enabling the creation of autonomous and intelligent manufacturing systems, with the ability to self-configure, self-monitor, or even self-repair, thereby increasing the efficiency and flexibility of manufacturing [9, 14].

With the development of the Internet of Things, there has also come the development of technologies that allow various devices to recognize and communicate with each other. One form of such communication is RFID (Radio-Frequency Identification), which makes it possible to read and sometimes write data on an integrated chip, which is the main component of the so-called RFID tag (Fig. 1).

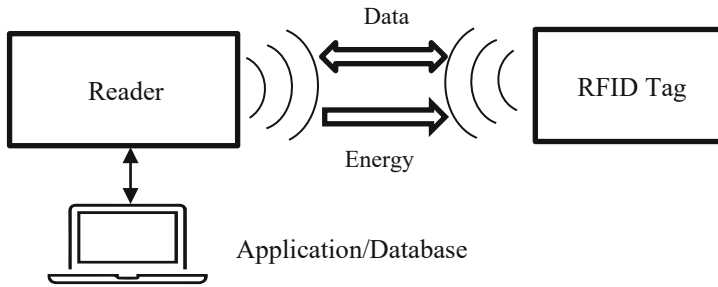


Fig. 1. Scheme of RFID technology [20]

RFID tags can take different forms and shapes depending on their intended use. The basic form of an RFID tag takes the form of a transparent plastic label, known as an inlay, to which an electronic chip and antenna are connected. A label, on the other hand, is an RFID tag usually made of paper, on which additional data, such as a barcode, can be printed. The RFID tag can also be placed in a special casing that protects the chip from external factors (impacts, temperature, humidity, water) [20].

There are several standards and varieties of RFID communication [13, 18, 20]:

- LF (Low Frequency) 125 kHz (Unique), close range (a few cm), data read-only, tag no. (EPC- Electronic Product Code, usually 96 bits),
- HF (High-frequency) 13.56 MHz (Mirafare, NFC - Near-field communication standard for short-range communication), short to medium range (several cm to 1 m), data writing and reading (ECP + user data 1–4 kbit, depending on the version),

- UHF (Ultra-High Frequency) 860 to 960 MHz – long-range (10 s to 100 s of meters), writing and reading a large amount of data (512 bits to 128 kbit).

The book [14] presents a review of the main research directions on the application of RFID technology in the Industry. The main area concerns inventory management systems [2], and object location detection [23]. RFID tags and technology have significant advantages, which include durability, small size, the ability to read and write data repeatedly, or to protect data from unauthorized access [16].

RFID also offers the advantage of not needing a separate power supply for the tag, which receives the energy it needs for operation directly from the electromagnetic field emitted by the reader. As a result, the operating range of this communication is limited, but the tags are very small and lightweight which allows them to be widely used in a variety of applications, including in commerce for the identification of products, containers/pallets, tools, employees, production tracking, warehouse inventory, room access control, time recording, etc. [3, 16, 23].

Unlike optically readable codes (QR, barcode), the RFID tag does not need to be visible to be read or written. Disadvantages of this technology include vulnerability to cyber-attacks (the possibility of remote copying of tag data) and destruction by a high-power electromagnetic impulse [3, 20].

The article [17] presents possible architectures and corresponding implementations in real-world scenarios, along with the most critical challenges affecting communication and sensing performance. In addition, RFID can act as a link between process flow data and physical asset data. IoT helps collect more data with RFID tags, readers, and sensors which improves transparency and efficiency in production management.

Research is also being conducted on the use of RFID technology in new areas including robotics. Article [22] presents applications of RFID (radio frequency identification) technology in combination with industrial robots. An Industry 4.0 scenario that makes workpieces smart by equipping them with RFID transponders was developed. RFID tags can store information in a decentralized manner, which can then be used to optimize the robot's operation. In the example, first, the width of the workpiece is identified by the robot's gripper and stored in the object. The value can then be read out and used to further catch the object quickly and sensitively. The article gives further suggestions for the kind of information that can be useful for storing robotic workpieces.

Another paper discusses an application of RFID-based part identification for performing robotic assembly operations in random mixing mode [15], using the integration of vision and RFID object identification [5] or using the properties of the RFID signal for “pre-touch” to facilitate grasping of various objects [4].

A paper [8] proposes a new paradigm for the interaction of robots with manipulation objects via radio frequency identification (RFID). Robots interact with physical objects, which in turn provides the robots with information showing how the objects should be manipulated, allowing robots to easily identify unknown objects and manipulate them.

One area of application for cyber-physical systems can be so-called “smart” grippers, which feature increased functionality and application flexibility due to integrated sensors and communication systems. An example of such a gripper could be the SCHUNK EGI intelligent and multifunctional gripper [10]. It is equipped with a network server with a Profinet interface, an active brake, and several sensors, which allow individual

programming of the stroke (up to 57.5 mm per jaw) and equally flexible gripping force values (in the range of up to 100 N). This allows the gripper to catch parts of different shapes, surfaces, and strengths. It can be used to gently and reliably handle flexible, deformable, or brittle parts [10].

3 Design of the Robot Workstation

The design of the Astorino teaching robot workstation was based on a brainstorming session conducted by students, during which the initial design of the workstation was determined. In the next stages of the project, concepts for the station's equipment were developed, which were modeled using CAD systems. The prototype components were made using a 3D printer. The final stage was the assembly of the instrumentation and integration of the systems into the workstation, as well as the programming and testing of the developed solutions. As a result, a prototype workstation was created, which is shown in Fig. 2.

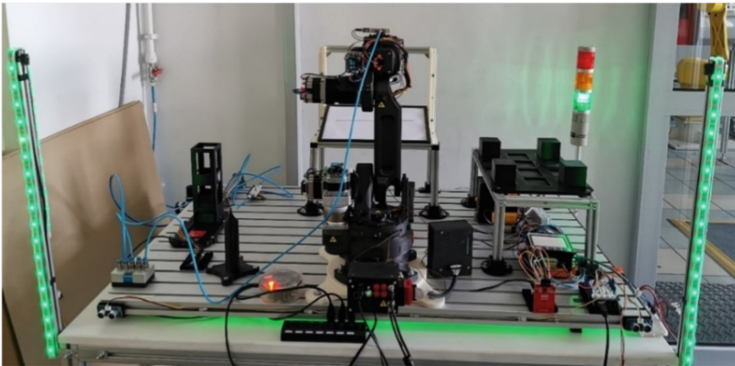


Fig. 2. Developed workstation with Astorino robot

The basic element of the workstation is a mobile platform, on which the robot was mounted, as well as equipment components, such as grippers, designed and made using 3D printing technology.

3.1 RFID Gripper Design

One of the stages of the PBL project was to equip the robot with a “smart” gripper that could work with the latest radio frequency identification (RFID) product identification systems, which are currently being used more widely in Industry 4.0. Taking into account the available technical capabilities, and considering that the Astorino robot used an Arduino microcontroller-based control system (Teensy 4.0) [21], it was decided that our system would be based on Arduino products. During the development of the various concepts in the team, we were keen to achieve a short distance of the reader concerning

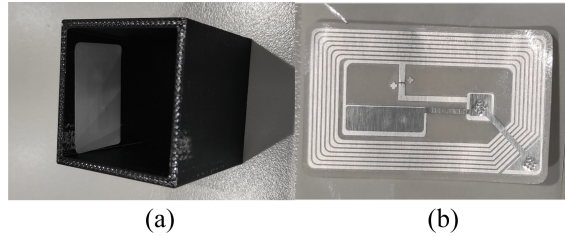


Fig. 3. a) Item with RFID tag, b) RFID sticker

the RFID tag sticker itself located on the manipulated object (Fig. 3) and to limit the thickness of the gripping jaws themselves.

Based on the standard Festo pneumatic gripper [6], new jaws were designed for the gripper to allow attachment of the RFID-RC522 Iduino sensor board [7]. A variant operating at the HF frequency of 13.56 MHz was chosen, which requires connection to an Arduino-type microcontroller for proper operation. The developed prototype of the gripper was adjusted to the size of the RFID board, which was $61.5 \times 40 \times 5$ mm. By making the gripping jaws thicker and maintaining the appropriate distances between the tips when the gripper closes, the grip range up to 52.8 mm was achieved. The aforementioned concept is shown in Fig. 4.

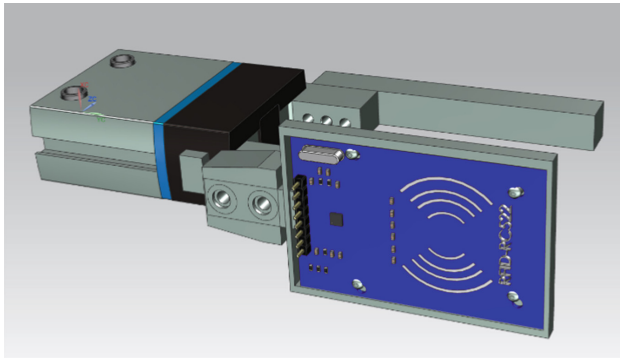


Fig. 4. CAD design of RFID gripper

The gripper prototype used a Festo pneumatic gripper [6] and gripping jaws printed on a FDM (Fused Deposition Modeling) 3D printer.

The thickening of the jaws did not affect the properties of the RFID reader. The mounting of the plate on one of the gripping arms was based on pressing in on four pins, which ensured a permanent fix.

An important aspect to consider when designing this component was the wiring of the RFID board. The connection pins had to be easily accessible and as close as possible to the gripper itself. The board could not go in too deep into the space in the jaw, because it would not be possible to connect the pins to the wires from the controller. The mounted gripper jaw with RFID reader is shown in Fig. 5a.

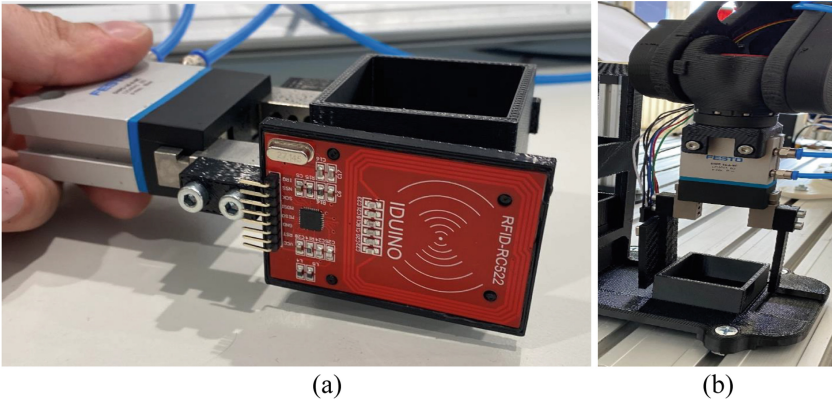


Fig. 5. a) RFID reader attachment in the gripper jaw, b) mounted on the robot

After mounting the sensor board and connecting it to the controller, the system passed tests on the Astorino robot workstation (Fig. 5b).

Signals from the RFID reader including, among other things, the product identification number are read by the Arduino microcontroller. To create a program for the microcontroller, open-source libraries available on the Internet were used, including source code that allows the use of the RFID reader along with instructions for connecting to the microcontroller [1]. Programming the application for the robotic station will be the subject of further development work.

4 Testing the Ability to Read Various RFID Tags

Since RFID technology has some technical limitations, additional tests were conducted involving the ability to read various RFID tags placed on manipulated objects, using a prototype of the developed gripper.

To test the ability to identify objects using RFID tags and the robot gripper, a series of objects in the shape of cubic boxes were used, which were printed on a 3D printer of the Stratasys FDM 360 type. Different tags were placed on the boxes with dimensions, 50×50 , 40×40 , 30×30 , 20×20 , and a test at what distance their reading is possible was conducted.

The RFID 13.56 MHz HF type labels shown in Table 1 were tested.

Table 1. Technical data of selected 13.56 MHz RFID labels.

No.	Name	View	Color	Dimensions [mm]	Comments
1	No name - white		White	Φ20	Mirafare Classic 1K 13.56MHz
2	Adafruit 360		White	Φ25	13.56MHz RFID/NFC White Tag - Classic 1K
3	Adafruit 361		Clear	Φ25	13.56MHz RFID/NFC Clear Tag - Classic 1K
4	Adafruit 362		White	25x40	13.56MHz RFID/NFC Sticker - Classic 1K
5	Sticker RFID		Black	Φ25	Waterproof sticker RFID MIFARE 1k
6	Sticker RFID for metal		Black & Blue	Φ35	Waterproof sticker NFC/MIFARE S50 Classic 1k
7	No name – silver		Silver	Φ25	Waterproof sticker RFID MIFARE 1k
8	Sticker NFC		Blue	Φ35	Sticker for metal NFC/MIFARE S50 Classic 1k
9	No name card		White	55x85	Mirafare Classic 1K 13.56MHz
10	Marker RFID		Black	75x25x5	Durable marker for metal with increased resistance to mechanical damage and moisture, dust, low temperatures MIFARE S50

The labels can only be read with the correct orientation of the reader and the labels, as they must be parallel to each other in close proximity. Due to the possibility of the gripper jaw damaging the labels, they were glued on the inside surface of the plastic boxes.

In the first step, the test stand was set up in the configuration shown in Fig. 6, where the gripper jaw with the reader positioned on the outer side catches the surface of the box with the label on the inner side.

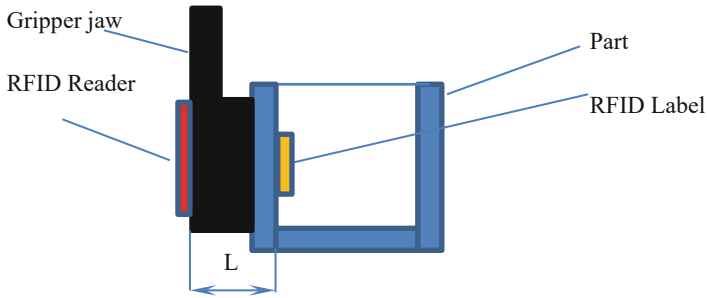


Fig. 6. Configuration of the stand in study 1

In this situation, the thickness of the object wall, which is 3 mm, and the thickness of the gripper jaw, which is 9 mm, affect the total distance L about 12 mm, between the reader and the label. During the test, it was possible to read data from all 10 labels. Since the labels are characterized by different sizes and the specifications give different reading ranges (or no data), another test was carried out with the bench configuration shown in Fig. 7. A series of cubic boxes with outer dimensions of 20, 30, 40, and 50 mm were used, and the labels were placed on the opposite side of the box with respect to the jaw with the reader.

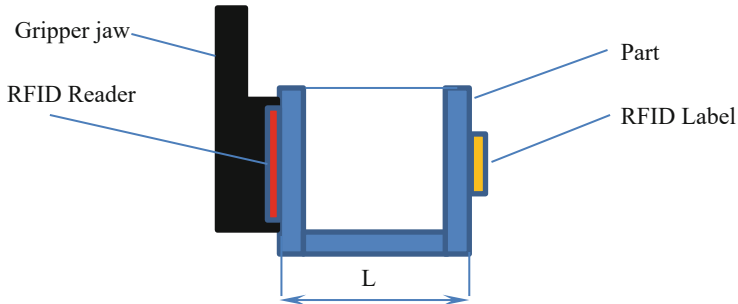


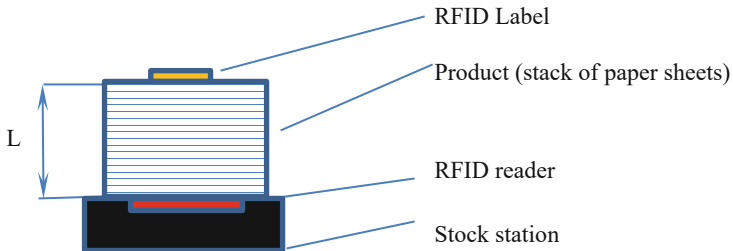
Fig. 7. Configuration of the stand in study 2

The results of the test (Table 2) showed that the labels differ significantly in the reading distance, which is close to the dimensions of the label and, more specifically, to the dimensions of the RFID antenna placed inside the label.

Table 2. Results of the RFID label reading distance test (Yes/No)

No	Name	Read distance RFID [mm]				
		12	20	30	40	50
1	No name - white	Y	Y	N	N	N
2	Adafruit 360	Y	Y	N	N	N
3	Adafruit 361	Y	Y	N	N	N
4	Adafruit 362	Y	Y	Y	Y	N
5	Sticker RFID	Y	Y	N	N	N
6	Sticker RFID for metal	Y	Y	N	N	N
7	No name – silver	Y	Y	Y	N	N
8	Sticker NFC	Y	Y	Y	N	N
9	No name card	Y	Y	Y	N	N
10	Marker RFID	Y	Y	N	N	N

Due to the significant differences in reading distances, one more test was conducted, using a stock station with a reader placed from below and a stack of paper cards of adjustable height. A schematic of the stand is shown in Fig. 8 and the results are in Table 3.

**Fig. 8.** Stand configuration in study 3

The results show that the tested labels have a small reading range of about 20–40 mm, which is related to the limitations of the used 13.56 MHz HF RFID technology and their label (chip) dimensions and materials used. This parameter is of immense importance in determining their applicability.

Table 3. Results of study 3

No	Name	Dimensions [mm]	Maximal reading distance [mm]
1	No name - white	Φ20	20
2	Adafruit 360	Φ25	24
3	Adafruit 361	Φ25	25
4	Adafruit 362	25x40	40
5	Sitcker RFID	Φ25	25
6	Sticker RFID for metal	Φ35	20
7	No name – silver	Φ25	30
8	Sticker NFC	Φ35	35
9	No name card	55 × 85	31
10	Marker RFID	75 × 25 × 5	21

5 Conclusions

Industry 4.0 is a multifaceted concept that integrates various technologies enabling the creation of autonomous and intelligent manufacturing systems with self-configuration and self-control capabilities to improve manufacturing efficiency and flexibility. One of the main areas of the Fourth Industrial Revolution is automation and robotics, as well as communication systems between various industrial objects such as machines, robots, or products, using cyber-physical systems which use RFID wireless communication technologies and the Internet of Things. Modern information technologies make it possible to change the previous paradigm of a centralized production system that mass-produces the same products to a so-called “smart product”, which can contain individually tailored information used to control the manufacturing process of a given product. The research is a first step in the development of equipping robots with RFID grippers and software of smart objects with RFID tags so that the information contained in the object allows flexible modification of the robot’s work program.

The results obtained so far show that the tested tags have a small reading range of about 20–40 mm, which is related to the limitations of the 13.56 MHz HF RFID technology used, as well as the dimensions of the tags and the materials used. Such values of the parameter are sufficient for basic applications, but are of great importance in determining the applicability of RFID labels.

The results of the study will be used for further development work on RFID grippers including use of professional sensors and modification of the software. Experiments on storing the information on RFID tags are also planned. This approach will enable the decentralized storage of information about items, while at the same time allowing the history of an item to be tracked through the manufacturing process.

Another research area is the integration of a vision system [19] and control of the robot arm using a tablet via the Internet [11], as well as an experimental study of robot and gripper vibrations [12].

The implementation of the PBL project enabled the students to acquire practical knowledge in the fields of automation, robotics and mechatronics while teaching teamwork, objective evaluation of the results obtained, and creative solutions to the problems encountered.

Acknowledgment. The research reported in this paper is the result of the PBL project co-financed by the European Union from the European Social Fund in the framework of the project "Silesian University of Technology as a Center of Modern Education based on research and innovation" POWR.03.05.00-00-Z098/17.

References

1. Arduino Project Hub. RFID-RC522 Sensor. <https://create.arduino.cc/projecthub/person87/rfid-rc522-sensor-7fc37b>. Accessed 12 June 2022
2. Bagchi, U., Guiffrida, A., O'Neill, L., Zeng, A., Hayya, J.: The effect of RFID on inventory management and control. In: Jung, H., Jeong, B., Chen, F.F. (eds.) *Trends in Supply Chain Design and Management*. Springer Series in Advanced Manufacturing, pp. 72–92. Springer, London (2007). https://doi.org/10.1007/978-1-84628-607-0_4
3. Bednarek, M., Dąbrowski, T.: Bezpieczeństwo transmisji danych w przemysłowym systemie sterowania. *Biuletyn WAT* **64**(4), 83–96 (2015). <https://doi.org/10.5604/12345865.1186229>
4. Boroushaki, T., Perper, I., Nachin, M., Rodriguez, A., Adib, F.: RFusion: robotic grasping via RF-visual sensing and learning. In: *Massachusetts Institute of Technology. SenSys '21: Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*, November 2021, pp. 192–205 (2021). <https://doi.org/10.1145/3485730.3485944>
5. Chong, N.Y., Tanie, K.: Object directive manipulation through RFID. In: *Society for Control and Robot Systems: Conference Proceedings 2003*. Vol. 10a, pp. 2731–2736 (2003). <https://koreascience.kr/article/CFKO200333239337897.pdf>
6. Chwytki równoległe FESTO DPS. https://www.festo.com/cat/pl_pl/products_DHPS?CurrentPartNo=1254045. Accessed 07 May 2022
7. Czytnik RFID RC522 13,56MHz SPI + karta i brelok - czerwony - Iduino ME138. <https://botland.com.pl/moduly-i-tag-i-rfid/>. Accessed 07 May 2022
8. Deyle, T., Tralie, C., Reynolds, M., Kemp, C.: In-hand radio frequency identification (RFID) for robotic manipulation. In: *Proceedings - IEEE International Conference on Robotics and Automation*, pp. 1234–1241. <https://doi.org/10.1109/ICRA.2013.6630729>
9. Fidali, M. (red.): *Przewodnik po technologiach Przemysłu 4.0*. Wydawnictwo Elamed, Katowice (2021)
10. Inteligentny chwytak do małych komponentów z aktywnym układem utrzymania siły chwytania. <https://automatykaonline.pl/Artykuly/Robotyka/Inteligentny-chwytk-dom-alych-komponentow-z-aktywnym-ukladem-utrzymania-sily-chwytnia>. Accessed 19 May 2022
11. Kaczmarek, W., Lotys, B., Borys, S., Laskowski, D., Lubkowski, P.: Controlling an industrial robot using a graphic tablet in offline and online mode. *Sensors* **21**, 2439 (2021). <https://doi.org/10.3390/s21072439>

12. Kaczmarek, W., Borys, S., Panasiuk, J., Siwek, M., Prusaczyk, P.: Experimental study of the vibrations of a roller shutter gripper. *Appl. Sci.* **12**, 9996 (2022). <https://doi.org/10.3390/app12199996>
13. Kamiński, A.: Inteligentna Fabryka – nowe trendy w rozwoju systemów informatycznych dla przemysłu, *Zarządzanie i Finanse J. Manag. Financ.* **16**(3/2/2018) (2018). http://www.wzr.ug.edu.pl/zif/11_9.pdf
14. Kanagachidambaresan, G.R., et al. (eds.): *Internet of Things for Industry 4.0*, EAI/Springer Innovations in Communication and Computing. https://doi.org/10.1007/978-3-030-32530-5_1345865.1186229
15. Makris, S., Michalos, G., Chryssolouris, G.: RFID driven robotic assembly for random mix manufacturing. *Robot. Comput.-Integr. Manuf.* **28**(3), 359–365 (2012). <https://doi.org/10.1016/j.rcim.2011.10.007>
16. Nath, B., Reynolds, F., Want, R.: RFID technology and applications. *IEEE Pervasive Comput.* **5**(1), 22–24 (2006). <https://doi.org/10.1109/MPRV.2006.13>
17. Occhiuzzi, C., Amendola, S., Nappi, S., D’Uva, N., Marrocco, G.: RFID technology for industry 4.0: architectures and challenges. In: 2019 IEEE International Conference on RFID Technology and Applications (RFID-TA), pp. 181–186 (2019). <https://doi.org/10.1109/RFID-TA.2019.8892049>
18. Platforma Przemysłu Przyszłości. <https://przemyslprzyszlosci.gov.pl/>, dostęp: 18 July 2022
19. Prusaczyk, P., Kaczmarek, W., Panasiuk, J., Besseghieur, K.: Integration of robotic arm and vision system with processing software using TCP/IP protocol in industrial sorting application. Paper presented at the AIP Conference Proceedings, vol. 2078 (2019). <https://doi.org/10.1063/1.5092035>
20. Radio-frequency identification. https://en.wikipedia.org/wiki/Radio-frequency_identification. Accessed 18 June 2022
21. Robot edukacyjny ASTORINO. ASTOR. <https://www.astorino.com.pl/>. Accessed 1 Dec 2022
22. Thormann, C., Winkler, A.: Localization and efficient grasping of objects by a manipulator using RFID technique. In: 2017 22nd International Conference on Methods and Models in Automation and Robotics (MMAR), pp. 670–675 (2017). <https://doi.org/10.1109/MMAR.2017.8046908>
23. Zhou, J., Shi, J.: RFID localization algorithms and applications—a review. *J. Intell. Manuf.* **20**, 695–707 (2009). <https://doi.org/10.1007/s10845-008-0158-5>



Development of an “Artificial Lung” System for Use in Indoor Air Quality Testing

Andrzej Kozyra¹ (✉), Aleksandra Lipczyńska², Piotr Koper², Radosław Babisz¹, Damian Madej¹, Konrad Nowakowski¹, Jakub Karwatka², and Dominik Tomczok²

¹ Department of Measurements and Control Systems, Faculty of Automatic Control, Electronics and Computer Science, Silesian University of Technology, Gliwice, Poland
Andrzej.Kozyra@polsl.pl

² Department of Heating, Ventilation and Dust Removal Technology, Faculty of Energy and Environmental Engineering, Silesian University of Technology, Gliwice, Poland

Abstract. “Artificial lung” is a device that simulates breathing process of occupants in a room. This allows you to safely test, e.g., the impact of HVAC systems on the spread of pathogens. The paper describes the concept of the device, its construction and performed tests. The project was implemented by students of two faculties: Automatic Control, Electronics and Computer Science, and Energy and Environmental Engineering. Project supervisors from both faculties provided substantive supervision and support for students.

1 Introduction

The thermal environment and indoor air quality in offices affect health, comfort and work efficiency [1]. Air distribution in a room is one of the main parameters determining the distribution of airborne pollution and the risk of cross-contamination [2]. The airflow in a mechanically ventilated room depends on the ventilation expenditure, the way of air supply and exhaust, and buoyancy forces caused by warm and cold surfaces in the room.

The most used air distribution system is mixing ventilation, designed to ensure an even concentration of pollutants throughout the room by diluting the polluted room air with clean supply air. However, the supply air is rarely thoroughly mixed with the room air [3]. As a result, zones with high concentrations of air pollutants, such as virus-containing aerosols, may occur.

The point at which the concentration of air pollutants is measured to determine users’ exposure is also important. Recent studies have shown that it is possible to achieve a significant difference in CO₂ concentration measured in the inhaled air compared to CO₂ concentration measured in the exhaust air and on the walls of a ventilated conference or office room [3]. These results show that further extended study on airborne pollution in reference to ventilation system design and operation is needed to achieve a productive and healthy indoor environment.

The impact of HVAC systems on the distribution of pollutants in a room is analyzed in laboratory conditions using tracer gases and an aerosol generator that simulates various types of pollutants [4]. The most crucial simulated contaminant is the pollutant emission

in the air exhaled by humans, which is the primary source of indoor contamination. COVID-19 has shown that airborne infections can cause not only dramatic loss of human life but also cause devastating economic and social disruption, including challenges to public health and people's well-being and jobs. Infectious respiratory pathogens such as SARS-CoV-2 are emitted in the aerosol produced by disease carriers when coughing, sneezing, talking and breathing [5].

Respiratory pathogens are simulated in laboratory conditions by dosing tracer gas (e.g., N_2O) or aerosols in "artificial lungs" into the air exhaled by one of the thermal manikins that simulate an infected person [6]. By measuring the concentration, e.g., in the air inhaled by other people simulated in the research space, the impact of the ventilation system on the risk of airborne infection can be realistically assessed. Such samples are collected in the inspired air system of "artificial lungs".

Artificial lungs can be used in the development and verification of the virtual model [7]. Digital twins are a virtual representation of the operation of a real object, in this case, the process of human breathing. The developed model can allow for a fully virtual simulation of the impact of human presence on the operation of the ventilation system.

The presented work results from a project carried out as Project Based Learning by students of the Faculty of Automatic Control, Electronics and Computer Science, and the Faculty of Energy and Environmental Engineering at the Silesian University of Technology under faculty supervision. The project aims to develop and implement a comprehensive "artificial lung" solution that allows for a realistic simulation of the human breathing process to emit pollutants in the exhaled air and to collect a sample of inhaled air for analysis.

2 "Artificial Lung" System Concept

The device should be able to adjust its operation to be able to simulate the breathing process for different levels of breathing. For example, while resting, a person takes about 17 breaths per minute (men about 19/min and women 16/min) [8]. The average tidal volume in a healthy person without hard work is about 500 ml for men and 400 ml for women [9]. The following equation can be used to implement the guidelines for controlling the respiratory system:

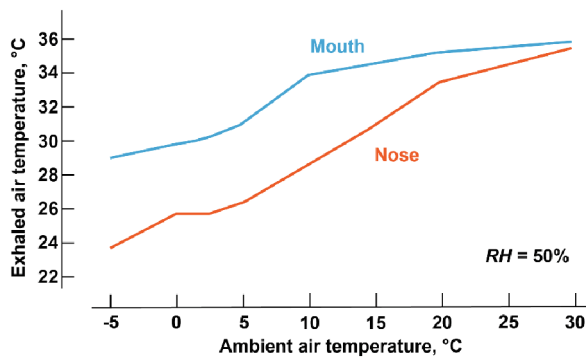
$$VT = f \cdot VE, \quad (1)$$

where: VT is the total minute volumetric flow expressed in mL/min, VE is the tidal volume in mL, and f is a respiratory rate (breaths/min). From the above analyses, it can be assumed that the average respiratory capacity may be 500 ml, and the average number of breaths 16/min. The respiratory cycle can be divided into four periods: inhalation, pause after inhalation, exhalation, and pause after exhalation. The entire cycle varies in time depending on, e.g., the activity performed and its difficulty level. The total breathing time decreases as the person's effort increases. Exemplary activities performed by a person presented in Table 1 were used to simulate breathing through the lung [10]. The simplified mode assumes work simulating light activity for which the pulmonary ventilation rate is 6 L/min, and the breathing cycle is 2.5 s – inhalation, 2.5 s – exhalation and 1 s – break [11, 12].

Table 1. Exemplary breathing patterns for adults [10]

Activity level	Inhalation time, s	Exhalation time, s	Break time, s	Tidal volume, ml
Rest	1.69	2.17	0.56	618
Light cycling	1.66	2.15	0.02	883
Heavy cycling	1.30	1.72	0.03	1354
Light mental work	1.32	2.38	0.3	590
Heavy mental work	1.24	2.61	0.26	585

The temperature of exhaled air depends on the exhalation method (through the nose or the mouth). Figure 1 shows how the exhaled air temperature changes in relation to room (ambient) air temperature at a constant relative humidity of 50% [12].

**Fig. 1.** Exhaled air temperature concerning ambient conditions and exhalation mode [12]

In the project’s first stage, a list of requirements and technical parameters of the device under construction was prepared. It was decided to develop an artificial lung system consisting of the following subsystems:

- 1) pulmonary ventilation system – inhalation,
- 2) pulmonary ventilation system – exhalation,
- 3) tracer gas dosing system,
- 4) exhaled air heating system.

Vacuum pumps with a flow range of up to 15 L/min were selected. Solenoid valves were selected to switch between inhalation and exhalation to simulate the selected breathing cycle. The correct temperature of the exhaled air will be maintained by heaters used in 3D printer systems and monitored by thermocouples. The Raspberry Pi platform was chosen to control the entire system.

3 Device Design

3.1 Pulmonary Ventilation Systems

Figures 2 and 3 show schematic diagrams of “artificial lungs” for inhalation and exhalation. Vacuum pumps provide adequate airflow to and from the thermal manikin. During the exhalation (Fig. 2), air containing the appropriate concentration of carbon dioxide and temperature is blown out of the dummy. In a thermally insulated container, the air temperature is regulated using a heater, and the tracer gas is dosed in so that the air pumped out of the device corresponds to the parameters of the air exhaled from the lungs. A mixing fan constantly stirs the air in the container. During exhalation, the air flows further through the flowmeter to the dummy’s mouth, where its temperature is additionally measured. At the moment when a breath is taken, in order not to stop the operation of the vacuum pump, the air is released outside with a bypass (outside the tested room).

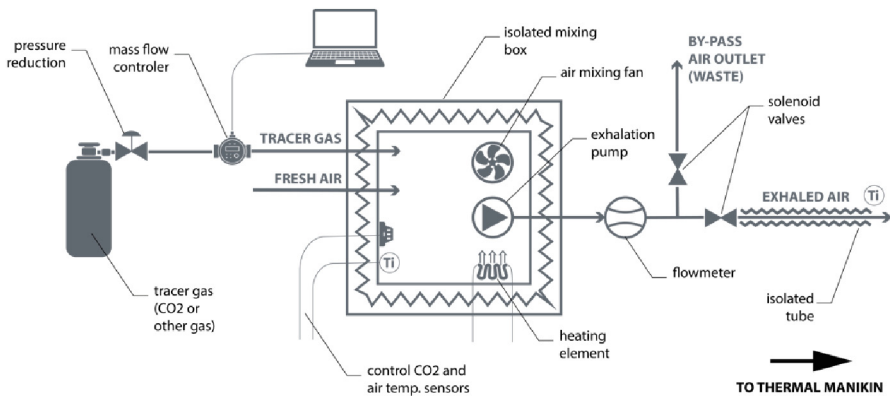


Fig. 2. Schematic diagram of the pulmonary exhalation system.

During inhalation (Fig. 3), the air from the tested room is sucked and inhaled air can be analyzed. The samples are taken from the sampling jar and sent to the gas analyser. The inhalation pump is working on a constant flow. During the exhalation cycle, the inhalation system is switched with solenoid valves not to take air from the tested room.

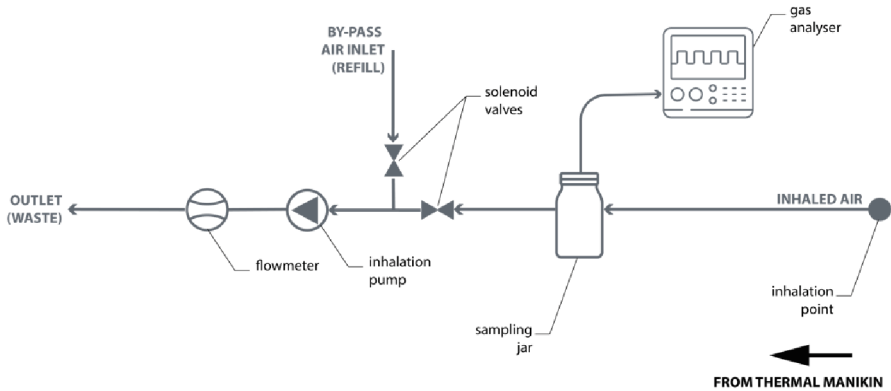


Fig. 3. Schematic diagram of the pulmonary inhalation system

3.2 Electronic System

The “artificial lung” system required the design of electronic circuits to control devices such as vacuum pumps, solenoid valves and heaters.

IRF520 MOSFETs with a maximum current of 5A to control solenoid valves and DFRobot Gravity modules with optical isolators and HM70P04K MOSFETs with a maximum current of 20 A to control pumps and a heater are used. In the case of pumps and heaters, MOSFETs are connected to the PWM outputs of the microprocessor, which enables smooth change of the heating power and speed of the pumps.

To control the basic parameters, several measurement systems were made to enable: (1) measurement of temperature and humidity (DHT22 sensor) in the exhaled air control chamber; (2) measurement of air temperature at the exhaled air outlet (thermocouple with Adafruit MAX31856 thermocouple amplifier); (3) measurement of carbon dioxide (MH-Z16 sensor) in the chamber; and (4) measurements of airflow in the inhalation and exhalation system (sensors YF-S402).

The main control unit is the Raspberry Pi 4B minicomputer with the Linux system installed. The used Raspberry Pi 4B version has 8 GB of RAM; two USB 3.0 and USB 2.0 connectors; two micro-HDMI connectors; a USB-C power connector; and a 32 GB microSD memory card. The minicomputer also has WiFi, Bluetooth, and an Ethernet port with a speed of up to 1000 Mb/s. The board also has 40 GPIO connectors, a CSI connector and a DSI connector. Raspberry Pi 4B can communicate with devices using UART, SPI, I2C and GPIO protocols. A microcontroller equipped in this way allows to control of devices via MOSFET, reads measurements of sensors with various interfaces and enables local control via a touch screen and remote access to measurement data via the network and SQL database.

Figure 4 shows the connection diagram of the microcontroller with the inhalation system, and Fig. 5 shows the connection system for the exhalation system.

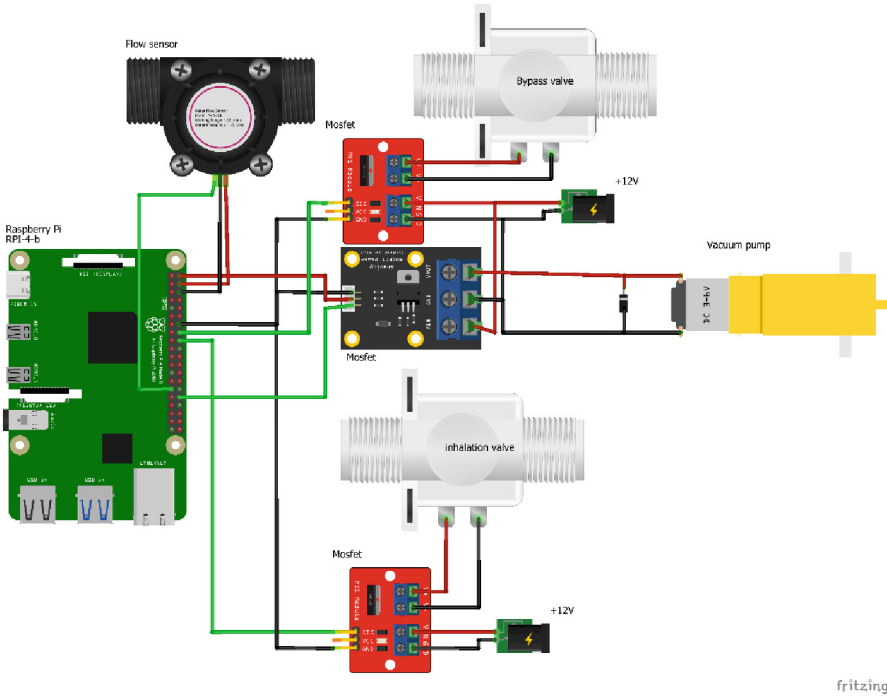


Fig. 4. Electrical diagram of the pulmonary inhalation system (image created with Fritzing).

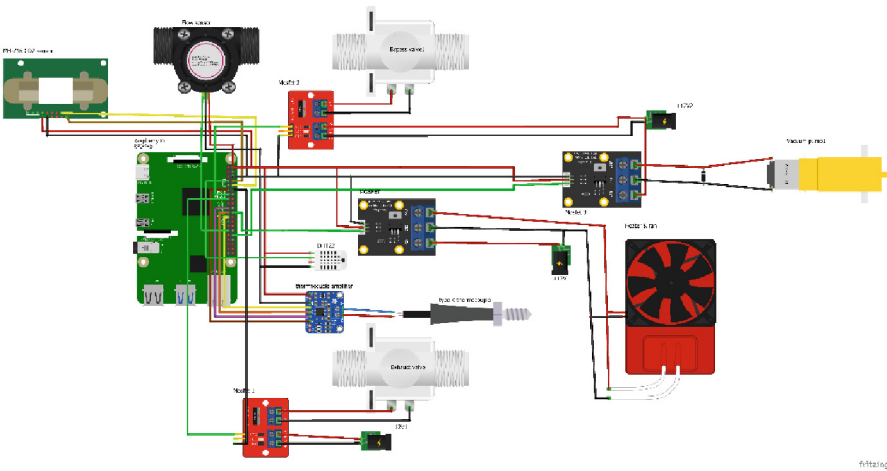


Fig. 5. Electrical diagram of the pulmonary exhalation system (image created with Fritzing).

3.3 Control Software

The “artificial lung” control program was written in Python. Several libraries were used to control all devices and sensors used in the project. The program is divided into tasks

controlling individual devices and supporting sensors. They are run at appropriate time intervals using the APScheduler library, allowing for smooth operation without blocking the program by delays in the written tasks serving individual elements.

These tasks were responsible for the following functions such as: (1) flow measurement based on the frequency of pulses from flowmeters; (2) temperature measurement using a thermocouple (SPI communication); (3) temperature measurement using the DHT22 sensor (1-wire communication); (4) measurement of carbon dioxide concentration using the MH-Z16 sensor (UART via USB communication); (5) control of vacuum pumps (PWM); (6) heater control (PWM); (7) control of solenoid valves via output pins; and (8) communication with the MySQL database.

The written program enables the archiving of all measurements using a MySQL database. Thanks to this solution, the user can observe the change of measurement parameters during operation. It also facilitates the detection and elimination of possible errors that may occur during the operation of “artificial lungs.” The database will also facilitate the development of a website and the visualization of the system’s operation on the Internet.

To make the operation of the “artificial lungs” easier, a touch screen was connected to the Raspberry Pi, with which it is possible to control the lungs using the graphical interface shown in Fig. 6. Several function buttons have been implemented that allow: (1) selection of one of the three preset operating modes or manual mode of “artificial lungs”; (2) stopping the operation of the “artificial lungs”; (3) closing the graphical interface; and (4) reading of the current temperature on the exhalation. Temperature monitoring allows you to assess the correct operation of the temperature control system of the exhaled air (Fig. 1). The temperature control algorithm turns on or off the heater in an insulated container depending on the temperature read by the thermocouple at the exhaust tube outlet. Additionally, the temperature control algorithm has protection against too high temperatures reached in the insulated container due to a carbon dioxide sensor that can operate in the range of up to 50 °C.

The ‘Close’ button closes the graphical user interface for operating the “artificial lungs.” The ‘Stop’ button stops the current operating mode of the lungs. The ‘Temperature reading’ button displays the temperature read by the thermocouple at the outlet tube, and the temperature is displayed above the ‘Close’ button.

There are start buttons to start the operation of the device for three predefined simulation models: “rest” (relaxation), “high activity” (heavy cycling), “normal” (light mental work), and manual in which the user sets parameters such as exhaled air temperature, inhalation and exhalation times, and airflow within specified ranges. The airflows for predefined modes were adopted based on Table 1.

Raspberry Pi with the display was placed in a special housing designed by the project team and printed on a 3D printer.

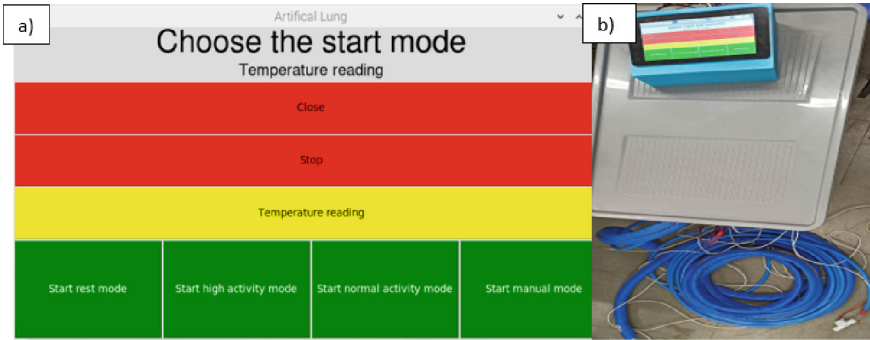


Fig. 6. The artificial lung: a) The graphical user interface, b) Completed device

4 System Start-Up and Tests

4.1 Calibration of Flowmeters

Turbine flowmeters with pulse output were used for the device, which had to be calibrated. An RTU-10-300 panel glass rotameter (accuracy class 2.5) was used as a reference. Pumps with a declared flow rate of 15 L/min were used in the project. However, preliminary tests have shown that the flow resistance is high. Only with two pumps is it possible to achieve the flow rate of a minimum of 12 L/min necessary to operate artificial lungs. Therefore, two pumps were also used for the calibration experiment. A series of measurements of the frequency of pulses generated by the flowmeter for various degrees of PWM duty cycle controlling the vacuum pumps were made (Fig. 7). The measurements carried out showed the maximum measurement error with the flowmeter used does not exceed 0.22 L/min.

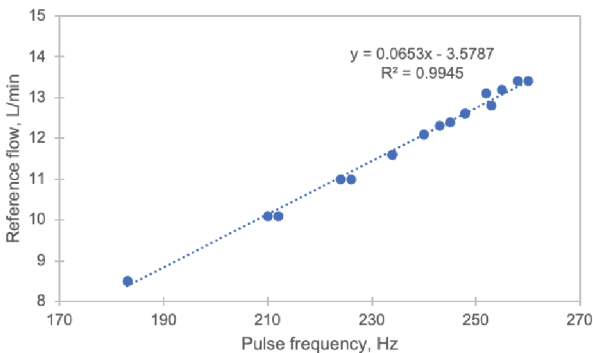


Fig. 7. Exemplary calibration for one of the flowmeters

4.2 Tests of the Exhaled Air Temperature Control System

The speed of reaching a steady state of air temperature in the mixing box was checked. Achieving a 45 °C temperature in an insulated container from 22 °C took approximately 25 min. Then, thanks to the temperature control algorithm, it was possible to maintain the temperature at the desired level with an error of ± 1 °C.

The exhaled air temperature is measured by a thermocouple placed inside the exhalation tube close to its end. An appropriate function controls this temperature. The “artificial lung” starts the correct work cycle when the exhaled air reaches the proper temperature.

A test of heating the air to the set temperature was carried out. The maximum temperature in the mixing box is limited to 45 °C by the operational parameters of installed devices. After two hours of keeping the box temperature at 45 °C, the exhaled air temperature only increased from 19.9 °C to 20.7 °C. This was because of a significant heat loss at the 7-m-long exhalation tube despite its isolation. It was decided to provide the exhalation tube with additional thermal insulation and wrap it with a heating cable to warm it along the entire length. The introduced changes enabled achieving the assumed temperature at the exhalation point. An additional control strategy will be developed to include the airflow changes within the exhalation tube due to the breathing cycle.

5 Summary

“Artificial lungs” have been constructed and tested. The following milestones were set and achieved during the project:

- Guidelines for the operation of “Artificial lungs” have been developed.
- A conceptual design of the device and control system has been developed.
- The developed device was constructed.
- Preliminary tests of the device have been carried out and necessary modifications have been made.
- Heating tests were carried out and a decision was made to add another heating element.

The tests showed that the device’s operating logic was correct, but the time needed to warm the exhaled air required to be shorter. It was, therefore, necessary to modify the air heating system.

The developed device will be used in the future in laboratory measurements concerning, e.g., the impact of ventilation air distribution, ventilation expenditure, and room arrangement on indoor air quality parameters, e.g., risk of infection.






The presented project also has a didactic aspect. Implementing the project in the form of PBL enabled students to acquire teamwork skills in a team of students from various faculties. Implementation of the project required knowledge in many fields: ventilation systems design, electronics, mechanics, 3D design, programming of real-time systems, databases, and many other skills.

References

1. Seppanen, O., Fisk, W.: Some quantitative relations between indoor environmental quality and work performance or health. *HVAC&R Res.* **12**, 957–973 (2006). <https://doi.org/10.1080/10789669.2006.10391446>
2. Lipczynska, A., Bivolarova, M.P., Guo, L., Kierat, W., Melikov, A.K.: Airborne infection probability in relation of room air distribution: an experimental investigation. In: *E3S Web Conference*, vol. 356, p. 05014 (2022). <https://doi.org/10.1051/e3sconf/202235605014>
3. Bivolarova, M.P., Markov, D., Snaselova, T., Melikov, A.K.: CO2 based ventilation control – importance of sensor positioning. In: *Roomvent & Ventilation Conference 2020, Torino* (2020)
4. Lipczynska, A., Kaczmarczyk, J., Melikov, A.K.: Thermal environment and air quality in office with personalized ventilation combined with chilled ceiling. *Build Environ.* **92**, 603–614 (2015). <https://doi.org/10.1016/j.buildenv.2015.05.035>
5. Buonanno, G., Morawska, L., Stabile, L.: Quantitative assessment of the risk of airborne transmission of SARS-CoV-2 infection: prospective and retrospective applications. *Environ. Int.* **145**, 106112 (2020). <https://doi.org/10.1016/j.envint.2020.106112>
6. Tang, J.W., et al.: Observing and quantifying airflows in the infection control of aerosol- and airborne-transmitted diseases: an overview of approaches. *J. Hosp. Infect.* **77**, 213–222 (2011). <https://doi.org/10.1016/j.jhin.2010.09.037>
7. Hosamo, H., Hosamo, M.H., Nielsen, H.K., Svennevig, P.R., Svidt, K.: Digital twin of HVAC system (HVACDT) for multiobjective optimization of energy consumption and thermal comfort based on BIM framework with ANN-MOGA. *Adv. Build. Energy Res.* **17**, 125–171 (2023). <https://doi.org/10.1080/17512549.2022.2136240>
8. Bendixen, H.H., Smith, G.M., Mead, J.: Pattern of ventilation in young adults. *J. Appl. Physiol.* **19**, 195–198 (1964). <https://doi.org/10.1152/jappl.1964.19.2.195>
9. Tobin, M.J., Chadha, T.S., Jenouri, G., Birch, S.J., Gazeroglu, H.B., Sackner, M.A.: Breathing patterns: I. Normal subjects. *CHEST* **84**, 202–205 (1983). [https://doi.org/10.1016/S0012-3692\(15\)33498-X](https://doi.org/10.1016/S0012-3692(15)33498-X)
10. Boiten, F.: Component analysis of task-related respiratory patterns. *Int. J. Psychophysiol.* **15**, 91–104 (1993). [https://doi.org/10.1016/0167-8760\(93\)90067-Y](https://doi.org/10.1016/0167-8760(93)90067-Y)
11. Melikov, A., Kaczmarczyk, J.: Measurement and prediction of indoor air quality using a breathing thermal manikin. *Indoor Air* **17**, 50–59 (2007). <https://doi.org/10.1111/j.1600-0668.2006.00451.x>
12. Höpffe, P.: Temperatures of expired air under varying climatic conditions. *Int. J. Biometeorol.* **25**, 127–132 (1981). <https://doi.org/10.1007/BF02184460>



Recent Advances in Artificial Autonomous Decision Systems and Their Applications

Andrzej M. J. Skulimowski^{1,2} , Inez Badecka^{1,2} , Masoud Karimi^{1,2} ,
Paweł Łydek¹ , and Przemysław Pukocz^{1,2} 

¹ Decision Sciences Laboratory, Chair of Automatic Control and Robotics, AGH University of Science and Technology, 30-059 Kraków, Poland

ams@agh.edu.pl

² International Center for Decision Sciences and Forecasting, Progress and Business Foundation, 30-048 Kraków, Poland

Abstract. This article presents the methodological background and an overview of recent applications of artificial autonomous decision systems (AADS), endowed with intelligent coordination algorithms. The research results comprise modelling the decision-making processes in autonomous anticipatory systems, specifically the decisions made in teams composed of robots and humans. These can be modelled as timed anticipatory networks. We present selected real-life applications of anticipatory decision support methods, such as industrial safety management systems and coordination of ground unmanned autonomous robot teams. New algorithms have been proposed to solve multicriteria combinatorial optimization problems that occur when selecting safety strategies, such as AADS evacuation from endangered areas. Decision analytics, based on reference sets, is proposed to solve optimal supervisory control problems related to coordinated autonomous robot deployment. Finally, we discuss the scenarios of future research on AADS and their applications in developing decision algorithms for teams of autonomous harvesting robots or robotic inspection of large industrial plants.

Keywords: artificial intelligent systems · autonomous robots · anticipatory networks · multicriteria decision support

1 Introduction to Artificial Intelligent Autonomous Systems

This paper refers to two main research areas of artificial intelligence (AI) applications in automatic control: intelligent decision support systems (iDSSs) and autonomous robotics. Both are linked by applications of novel research tools, based on multi-criteria analysis, machine learning and decision algorithms for autonomous systems. Similar methods have also been applied to modelling long-term business cooperation and related decision processes [16]. Real-life applications are supported by diverse implementations of decision support engines dedicated to the coordination of multiple autonomous robots [19], optimization of investment decisions, multicriteria evaluation of innovative project impact [13], and multicriteria control of water reservoir systems.

The (intelligent) *Artificial Autonomous Decision Systems* (AADSs) have been described in [14] as systems capable of making decisions at one of the freedom of choice (freewill) levels defined in [12]. This AI-inspired notion arose during an analysis of decision-making processes in autonomous inspection robots and web crawlers [14]. The AADS are classified by assigning them to one of four classes of freewill-based autonomy, defined within the conceptual apparatus of multicriteria decision theory [12]. According to this theory, the following three classes are distinguished:

- Systems capable of freely choosing decisions from a given set of alternatives.
- Systems that can independently expand the scope of their decisions (remove, move or release constraints).
- Systems capable of changing the purpose of their action, the goals to be reached, or any formally defined optimization criteria.

The fourth class are systems endowed with artificial creativity [12]. This classification allows for a systematic study of autonomous robots that is presented further in this article. AADS is also a background notion for various real-life applications in robotics, iDSS design, autonomous web crawlers and in many other areas.

Another relevant concept for the in-depth study of autonomous systems is that of the *anticipatory system* [10]. By definition, a system S is anticipatory if it is capable of building a future model of itself and of its environment and then use them when planning its own actions or making other decisions. This notion, used first in systems biology, has recently been extended to anticipatory networks (ANs) [13], linking different anticipatory systems as nodes. Further information about ANs with references to relevant papers can be found in Sect. 2.2.

The AADS-related research topics presented in this article have been selected while taking into account real-life industrial needs and synergies with research projects within the field of autonomous systems, conducted by the cooperating teams. These include simulation of inspection [15] and harvesting robots [19], group decision analysis in web communities modelled by anticipatory networks, as well as iDSSs used in industrial safety management [17], smart agriculture, technological roadmapping and strategic decision making [16]. It is worth noting that the latest research findings presented in this article have been continually included in the curricula at all levels of study in the field of Automatic Control and Robotics. Following this, members of student scientific societies and PhD students have shown a keen interest in conducting research on the above topics.

A schematic representation of the coverage of topics presented in this article is depicted in Fig. 1. The theoretical aspects are presented primarily in Sect. 2, while applications are discussed in Sects. 3 and 4. Specifically, Sect. 2 presents research related to anticipatory models of autonomous decision processes. Section 3 is concerned with iDSSs designed to coordinate the evacuation of equipment in industrial safety systems, both autonomous and human-supervised. Section 4 presents an overview of further research directions, a discussion as well as conclusions.

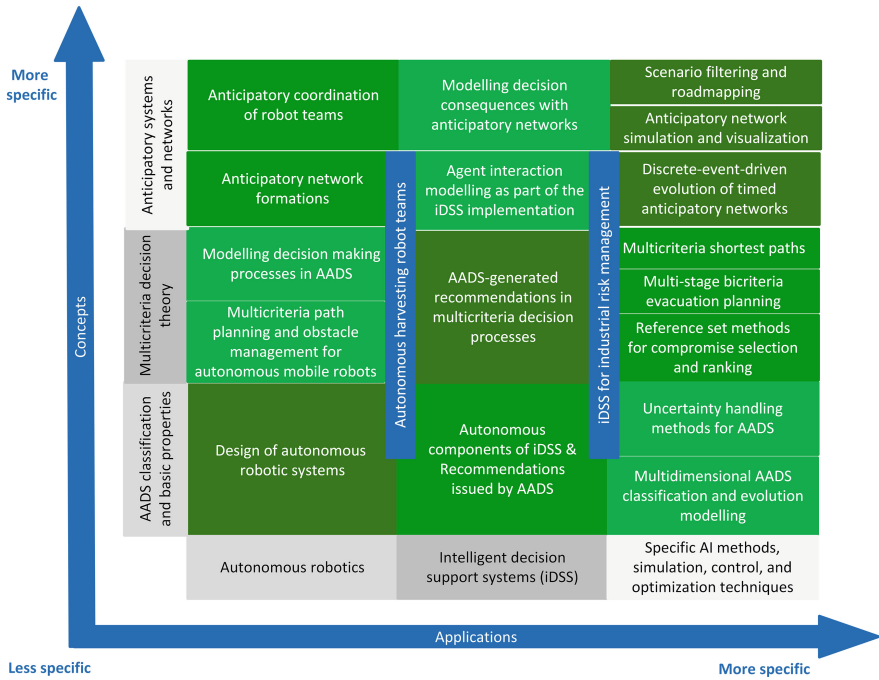


Fig. 1. Structure of topics presented in this article. Two blue vertical bars denote the currently studied real-life application areas and their relation to the background topics.

2 Decision Analysis in Autonomous Systems

One of the primary motivations for the research on AADSs was the multilevel analysis of the consequences of decisions in multi-stage planning processes [11]. Inspired further by the anticipatory systems theory of Rosen [10], the influence of anticipatory preference theory increased. It resulted in extending the scope of applications of this theory to multicriteria decision problem solving in iDSSs as well as in the modelling of autonomous robots.

In the decision processes within the stochastic environment with random controls and disturbances, the analysis of consequences can be modelled by causally dependent random variables embedded in Bayesian networks [4, 21]. Additional approaches used in autonomous robotics employ discrete event systems [15], where Markov decision processes [3] are commonly used to identify optimal decisions. However, for predominantly deterministic planning of autonomous service robot activities, the above-mentioned stochastic decision models proved insufficient. Including stochastic uncertainty in the ANs provided models of decision consequences, based on stochastic forecasts and vector utility functions [11] in causally related multicriteria optimization problems [13].

Modal logic is another approach increasingly used to classify and design autonomous systems. It is particularly relevant for designing anticipatory multi-agent models. Among

modal logics used in autonomous robotics, the most popular are the BDI (*Belief-Desire-Intention*) [2] and the *dynamic epistemic logic* (DEL, [1]). Together with different temporal logics, these two modal logics are also suitable for analyzing agent decisions in anticipatory systems and networks.

In Sect. 2.1 below, we outline the multicriteria optimization background and decision-theoretical foundations of AADSs.

2.1 Multicriteria Analysis Background of Decision Autonomy

The underlying multicriteria problems to be solved at different times, corresponding to decisions modelled by an anticipatory network, can be formulated as:

$$(F = (F_1, \dots, F_N) : U \rightarrow E) \rightarrow \min(\Omega), \quad (1)$$

where $F = (F_1, \dots, F_N)$ is a vector criterion, $U \subset \mathbb{R}^k$ is a set of potentially acceptable decisions, $E \subset \mathbb{R}^N$, and $\Omega : E \rightarrow 2^E$ is a certain preference structure that introduces the partial order $<_{\Omega}$ in E according to the formula

$$x <_{\Omega} y \Leftrightarrow^{\text{df}} x \in \Omega(y) \quad (2)$$

Definition 1. The set of *nondominated decisions* in U with respect to problem (1) is defined as.

$$P(U, F, \Omega) := \{u \in U : \forall v \in U [F(v) \in \Omega(F(u)) \text{ implies that } F(v) = F(u)]\}. \quad (3)$$

The set of non-dominated ratings, $FP(U, \Omega)$, is defined respectively as

$$FP(U, \Omega) := F(P(U, F, \Omega)) \quad \blacksquare$$

Usually, the sets $\Omega(y)$ are convex cones in \mathbb{R}^N and are constant. Decision makers that select decisions from the nondominated subset (3) are termed *rational*.

It follows from Definition 1 that two nondominated decisions are never mutually comparable with respect to the partial order $<_{\Omega}$ in U , so additional preference information must be used to select a compromise decision from the set $P(U, F, \Omega)$. Among various plausible models of decision-maker preferences, the anticipatory preference structure, based on a model of the future states of environment, proved particularly useful in solving multicriteria planning problems. Incorporating further preference information in (1)–(2) is equivalent to adjusting the preference structure Ω , according to updated knowledge about the dominated and dominating decisions for some $y \in U$. This update procedure is termed *preference modelling* or multicriteria decision making process [13]. Such a process is termed *inclusive* if the set of decisions comparable with a given $y \in U$ expands with subsequent process steps, i.e., if all former preference relations of type (2) associated to the process

$$\Omega_0 \rightarrow \Omega_1 \rightarrow \dots \rightarrow \Omega_n \quad (4)$$

satisfy the implication

$$i \leq j \Rightarrow [x <_{\Omega_i} y \Rightarrow x <_{\Omega_j} y] \text{ for } 1 \leq i \leq j \leq n, \quad (5)$$

where $\Omega := \Omega_0$. It is easy to see that the property (5) results in contracting the subset of nondominated decisions for subsequent stages of the process. Likewise, when the previously expressed preferences are preserved during the process, a rational decision maker needs to survey a smaller set when looking for a compromise decision. Ultimately, the set $P(U, F, \Omega_m)$ may consist of a single element and the decision process stops. This process may also be halted if the contraction is deemed sufficient to select a decision otherwise outside the process, e.g. randomly.

2.2 The Principles of Anticipatory Decision Processes

From now on, we will assume that the current and future environment of problem (1) is modelled as a network of interdependent agents capable of making decisions autonomously, hereinafter referred to as *decision makers*. Those that behave in a rational manner i.e., optimize a certain set of criteria according to a preference structure and select their decisions from $P(U, F, \Omega)$, are termed *multicriteria optimizers*. Their preference structures may be either explicit or fully unknown or partly unknown and discovered during the decision process. Optimizers capable of modelling themselves and forecasting the future states of their environment fulfil the definition of the anticipatory system. Multiple anticipatory optimizers can be embedded in an anticipatory network, which is formally defined as follows:

Definition 2. An anticipatory network A is a directed multigraph, $A = (Y, r, g)$, where the nodes Y are anticipatory optimizers and r is the digraph of an acyclic causal relation that describes the impacts of some anticipatory optimizers on other nodes of this digraph. The second digraph g is the graph of an acyclic relation termed *anticipatory feedback* that points out the preferences of some nodes as regards future decisions to be made by other agents represented in the network. In addition, if $(x, y) \in g$ then $(y, x) \in R$, where R is the transitive closure of r . ■

From Definition 2, it follows that the basic feature of anticipatory decision models is the possibility of indicating a certain subset of the network of optimizers with known – perhaps approximate – parameters of optimized problems i.e. criteria, constraints and preference structures. The nodes of this subnetwork are called *predictable optimizers*. The set of predictable optimizers is denoted by A_P . The prediction horizon of A_P is defined as the most distant period of decision making by nodes A_P . The selection of a compromise solution from nondominated decisions of a node z can be based on the forecasted solutions of predictable optimizers following z , according to the principle that the anticipated decisions be made at the nodes linked with z , as starting nodes of g fulfil the preferences indicated by z . An assessment of a given decision is made at the time of elaborating the forecast, taking into account the forecasted consequences of this decision in the network of dependent problems, indicated by the relation g . In the diagram shown in Fig. 2 below, O is the optimizer, X is the approximating operation of the set $P(U, F, \Omega)$, and

$$\psi : X(U, F, \Omega) \times A \rightarrow P(U, F, \Omega)$$

is the function selecting a nondominated decision to be considered as the best-compromise solution by the decision maker.

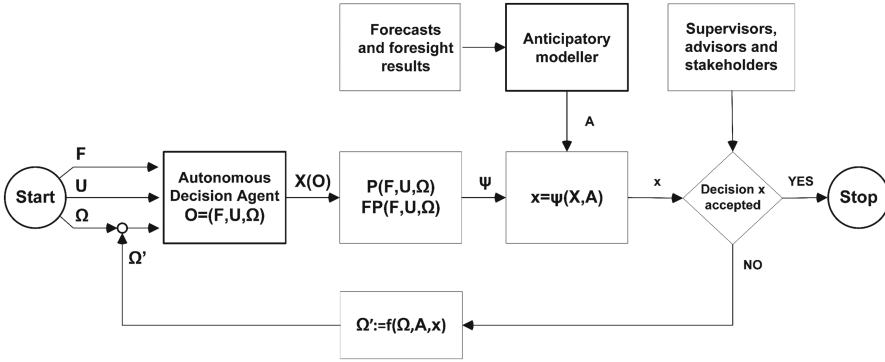


Fig. 2. The procedure of selecting a compromise solution in problem (1) with additional preference information provided by an anticipatory modeller in the form of an anticipatory network A modelling the present and future environment of the decision process.

In the anticipatory decision-making model presented in [11, 13], the parameters depend on the chosen decision within the current problem being solved. In addition to static optimization problems of type (1), one can also consider the situation when a set of predictable optimizers forms an *anticipatory control system* [19] for a multi-criteria optimal control problem with feedback:

$$\begin{aligned} \dot{x} &= h(x, u), \\ u(t) &= q(x(t), \hat{x}(t, t + \tau)), \end{aligned} \tag{6}$$

where $\hat{x}(t, t + \tau)$ is the prediction of the state of x after time τ available at time t , and the functions h and q respectively describe the dynamics of the open-loop system and the feedback. Similarly, anticipatory decision-making systems can be defined with discrete time.

The above theory can be applied to modelling and forecasting the consequences of decisions in causally dependent anticipatory systems [13, 19], where the ranges of future decisions are predicted based on assumptions of decision-maker rationality. The decision maker’s knowledge of the parameters of the predictable optimizer network, up to the forecasting horizon $t + \tau$, is used to make a decision at time t . It is assumed that the decision maker in problem (1) knows the characteristics of causal relations with predictable optimizers, the structure of the A_p , and how the preference structures are included in the decision algorithms of the A_p nodes.

Articles [13, 15] present some practical problems, where the influence on the sets of admissible solutions to future problems allows for the selection of a rational decision in the problem being solved. It is assumed that the scope of the decision in problem O_i depends on the solution of at least one of preceding problems i.e. O_{i-1} , and multivalued mappings $\varphi(i)$ are known such that $\varphi(i) := Y(i) \circ F_{i-1}$ for some multivalued function $Y(i) : F_{i-1}(U_{i-1}) \rightarrow \rightarrow U_i$, where “ $\rightarrow \rightarrow$ ” denotes a multivalued correspondence into U_i . $Y(i)$ describes the dependence of the decision scope in problem O_i on the values of criteria F_{i-1} that characterize the solution to problem O_{i-1} . In addition, it has been assumed that the anticipatory decision making principle is used by all predictable

optimizers that only take into account the consequences of their own decisions. In this model, multivalued functions $Y(i)$ describe causal relationships, and the consequence of choosing a decision in problem O_k consists of determining the range of solutions in one or more future problems O_m . To sum up, the decision maker solving the problem at O_k takes into account the native preference structure P_k , as well as preferences regarding the solutions to the future problem O_m . This principle exemplifies the above defined (cf. Definition 2) *anticipatory feedback* between O_m and O_k .

An example of how to relate decision problems in the anticipatory network is shown in Fig. 3.

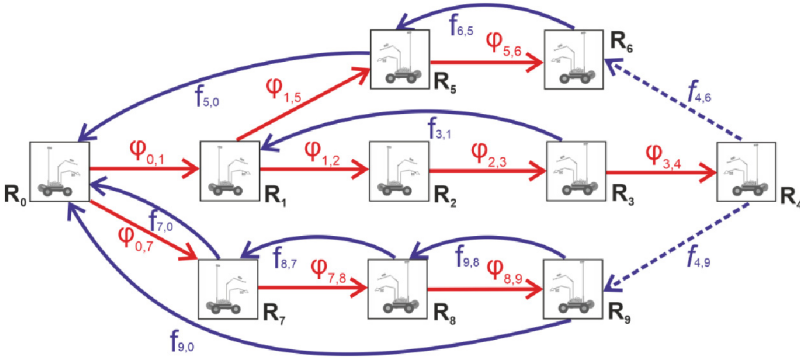


Fig. 3. Example of a causal network of decision problems solved by autonomous robots R_i , $i = 0, 1, \dots, 9$ (red edges) with anticipatory feedbacks (blue edges). The dashed lines indicate anticipatory feedbacks between causally unrelated agents, therefore they are irrelevant.

For the model described above, many variants of anticipatory decision-making problems can be formulated, differing in the way that consequences of previous decisions are taken into account and in the preference models. In the models of decision making in ANs considered thus far, these relations usually consisted of indicating a subset of decisions $V_{m,k} \subset U_m$, preferred by O_k . Constructive algorithms for solving anticipatory decision making problems in a network of predictable optimizers, where the set of decision alternatives is finite and all decision makers make rational decisions, can be found in [13, 15]. As previously indicated, rationality refers to the choice of a solution that is consistent with the decision maker's preference structure. Examples of solutions to anticipatory discrete problems and possible applications of this model to the construction of scenario forecasts of information technology development are presented in [13].

3 Decision Support Systems for Industrial Safety Management

As previously mentioned in Sect. 1, an important application area of autonomous systems and iDSS software architectures is the concept development and implementation of a decision support system for industrial risk management (IRM DSS, [17]). Here, we present a system capable of optimally exploring the advantages offered by the recent progress in autonomous or semi-autonomous systems, multi-criteria decision theory, signal analysis and fusion methods, and machine learning (ML). An implementation of

this class of system can significantly increase the level of technical and human safety in production enterprises in various sectors. The case study outlined below refers to an energy sector company (the Czatkowice Limestone Mine, Tauron Group), hereinafter referred to as CLM.

Management of technologies, based on AI methods, is part of the holistic security system of the enterprise. The data processing engine in the IRM DSS for CLM combines hitherto used photogrammetry and video monitoring techniques, as well as new solutions, such as radar systems supported by autonomous flying units (UAVs, drones) and ground inspection robots. The capabilities of modern active observation systems in visible, infrared and multispectral frequencies, as well as the use of radar, lidar, ultrasound, and other sensors, allowed the company to design a network of mutually supportive observation and measurement points, covering a vast area that is difficult to patrol. The concept of industrial risk management assumes an integration of the observation system with AI-based systems of identification and prevention of threats. The use of ML and other AI methods ensures that a harmless event, such as the appearance of a bird in an inspection camera view range, is distinguished from relevant hazards. The identification and further tracking of hazards by the system is necessary to ensure the required level of security. Such threats include the appearance of unauthorized persons, smoke or fire in the monitored area. Risk analyses carried out by CLM clearly indicate the need to protect the boundaries of the exploitation area against the threat of intrusion. Hazard detection and classification initiates an alert procedure.

The method of exploitation and the type of limestone deposit in CLM generates the risk of dangerous phenomena for people and machines, such as landslides of rock masses and rock falls. Due to the nature of the business, i.e. work on a continuously expanding vast area that is difficult or impossible to fence, the concept of the safety management system presented in [17] assumed the deployment of autonomous ground inspection robots or aerial vehicles (drones). The robots were equipped with appropriate measuring equipment for periodic analysis of the condition of the mining excavation, in order to identify potential threats. The recorded images from the inspection area are analyzed by an automatic image interpretation module using AI algorithms, based on artificial neural networks. This system component can indicate sites at risk of landslides and locate areas with a heterogeneous geological structure, which may naturally be displaced as a result of machine movement, or mining and shooting works that are carried out at the site. For this purpose, PTZ (pan-tilt-zoom) video surveillance cameras can also be used and manually directed by the operator to any part of the deposit that warrants in-depth observation.

Furthermore, the use of image analysis methods based on ML, in particular deep learning with convolutionary neural networks, makes it possible to exploit the potential of new technologies, and for purposes of volumetric estimation of excavated material. The data processing subsystem of an IRM DSS includes algorithms for real time filtering and fusing information received from various sources.

The development of optimal autonomous decision algorithms, based on processing all the above information, protecting privacy, and free of AI perils, is a relevant challenge [5, 17]. An application of anticipatory decision principles in the management of threat mitigating activities turned out to be an appropriate solution, by ensuring that the

system's autonomy remains within the confines defined by AN principles. Depending on the situation, decisions may be autonomously made or the system may be limited to only supporting a human supervisor's decisions. Due to the amount of data gathered in real time, it is necessary to implement modern data and decision analytics methods in an iDSS dedicated to solving predictive maintenance [9] and industrial security problems. Designing such a system requires a holistic approach, i.e. simultaneous implementation of modern ML-supported monitoring and prevention technologies and AI-based predictive and prescriptive analytics, ultimately aimed at building the resilience of the enterprise [23].

In the decision support system for industrial safety management in CLM, an important role will be played by algorithms for searching and prioritizing optimal evacuation routes of equipment from the endangered part of the enterprise area, and algorithms based on AI methods. They will make it possible to adjust the evacuation process to specific geological conditions of the exploited deposit in situations of danger, such as heavy rainfall. The equipment evacuation problem for n machines can be formulated as

$$\begin{aligned} \Sigma_{1 \leq i \leq n} d(M_i) + \Sigma_{1 \leq j \leq m} c(\alpha_j) &\rightarrow \min, \\ \Sigma_{1 \leq j \leq m} h(\alpha_j) &\rightarrow \min, \text{ s.t. } \alpha_i \in Q, \end{aligned} \quad (7)$$

where $d(M_i)$ is the value of damage that machine M_i sustained before it reached a safe site, Q is the set of admissible evacuation-related activities or coordinating decisions α_j . We assume that a selection of m of them will be undertaken by human rescue teams and by the autonomous machines independently. The cost of α_j , $c(\alpha_j)$, and the risk to human health related to activity α_j , $h(\alpha_j)$ are to be minimized.

Admissible activities Q include traversing obstacles [7] instead of avoiding them, when the impact of the risks of staying in a dangerous area is higher than the potential damage when crossing an obstacle. Specifically, when evacuating vehicles from a hazardous area, the problem of avoiding obstacles caused by landslides or falling rocks may arise. If the vehicles are moving on separate roads, which can be either passable or withdrawn from traffic due to the appearance of an obstacle, the evacuation problem then includes *combinatorial obstacle avoidance or crossing*. As another alternative activity in Q , some obstacles can be removed by evacuating equipment crews or rescue teams.

The solution to this problem consists of analyzing multicriteria shortest paths in a time-dependent graph of road links, where road sections (edges) with non-passable obstacles are ignored during the computation and re-included when obstacles are removed. The edges can be labelled with multiple parameters denoting expected energy consumption, time, distance, probability of damage when moving on this edge etc. The set of nodes contains the subsets of starting and destination nodes, the latter unique to each robot. Thus, the solution to problem (7) takes into account the constraints imposed by the multicriteria shortest path deconfliction [6] and the capacities of safe sites which are alternative evacuation targets. Such problems can also be solved by path prioritization or using the so-called bump-surface method [22], which can produce the exact solution of the shortest path problem. Real-time prescriptive analytics [20] can facilitate the decision choice by rescue teams.

Instead of optimizing criteria (7), one can also consider the sum of the expected arrival times of each evacuated machine to a safe area as a pre-criterion, which under some

natural assumptions, is correlated with the total value of losses, due to the equipment damage during evacuation. Genetic algorithms with search memory have been used to solve problems of this type, as well as problems arising in the selection of the iDSS configuration for industrial safety management [17].

An example illustrating the situation where evacuation algorithms, based on multi-criteria decision analysis, can be used to find best-compromise evacuation plan is shown in Fig. 4. Each alternative route of a machine to a safe area C_i is a decision characterized by the expected time to reach C_i , the expected damage and risks to the staff health. The system indicates the target area for each machine, while minimizing the summary values of the expected time to reach these areas, and damage during evacuation. When selecting the routes, the evacuation coordinator takes into account the maximum capacity of each safe area as constraints.

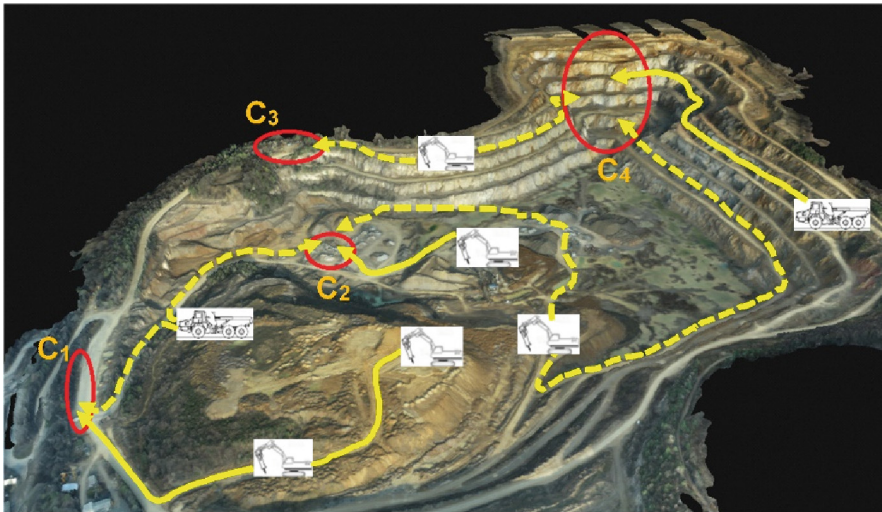


Fig. 4. An aerial view of a hazardous industrial area with ellipses marking safe areas C_1, \dots, C_4 as destinations for evacuation of equipment, supervised by a decision support system. The maximum parking capacity of each safe area C_i is c_i machines for $i = 1, \dots, 4$. Solid lines are routes that must be chosen by a machine attached to them, while dotted lines denote alternative routes. It is easy to see that if $c_i = 2$ for all i then the above example problem admits a unique solution. The area of the site covered by the photo is about 1.3 km^2 .

The simulation of various evacuation analysis and obstacle management algorithms that can be potentially implemented at CLM will be the basis for planning future investments aimed at reducing industrial risk, with a better organization of evacuation in case of an emergency. The benefits to the company of implementing an intelligent IRM DSS also include meeting increasingly stringent technical safety standards and providing employees with the highest level of safety.

4 Discussion and Conclusions

The above presented classification of intelligent autonomous systems and their applications is contingent on their ability to make decisions autonomously. For the AADS applied as virtual IRM DSS components, an equally relevant dimension of autonomy is the freedom to collect and process environmental parameters and information about the potential scope and availability of decisions. This notion is complementary to the former one and can equally be used to classify mobile AADSs. For example, a superior group, which meets both conditions of autonomy, includes artificial creative systems that are able to formulate and solve decision-making problems outside the previously known area [15], such as emerging creative scientific robots. This class of AADS is also the subject of intensive study within the context of autonomous harvesting robots [19].

The anticipatory network concept [13] made it possible to enhance the predictive and prescriptive analytics, with anticipatory decision modelling approaches, which can be used both in robotic AADSs and iDSSs.

Further research plans are aimed at developing the above-presented chapters of multicriteria decision and autonomous systems theory, as well as to extend the application areas of AADSs. This research will focus on the following topics:

- Methods for selecting compromise decisions in multi-level and hierarchical optimization and multicriteria ranking problems. A common feature of them is the possibility of simultaneous use of both direct preference information that concerns the values of criteria functions and dual information that concerns the substitution coefficients of these functions, in addition to a convex cone introducing a partial order in the criteria space. These methods can be used in technological and cooperative roadmapping [16], as well as in modelling the technological evolution of autonomous systems.
- Further extensions of the reference-point-based multicriteria decision making methods, for multicriteria model predictive optimal control problems, with a time-dependent reference set [12]. A multivalued function can be used instead of a reference trajectory to better model the uncertainty about the termination time of the process and the desired behavior of the system. This method allows adaptive fitting of the preference model to the updated predictions of the control system parameters in autonomous robot team action planning.
- The creation of qualitatively new models of strategic planning, which should lead to the development of significantly better strategic documents, e.g. those concerning the development of digital innovation. These models are based on the combination of dynamic programming for discrete-event system state transitions, modelled with variable structure networks, and on the generation of elementary scenarios in Markov Decision Processes [3, 8].

Research will also be pursued in the area of further implementation of multi-criteria analysis methods in decision support systems for crisis management purposes, including systems for managing simultaneous flood, fire, landslide and rock avalanche risks [17] that may co-exist with anthropogenic threats. It is expected that IRM DSSs will be increasingly supported by real-time drone- and ground-based mobile robot inspection. The related applications, such as autonomous energy supply for drone swarms inspecting large industrial plants are already under development.

In addition to the above, the foundations of the theory of autonomous systems and the theory of freedom of choice remain a challenging research area, as well as their applications to the design of robot cooperation [15]. These studies include, inter alia, the altruistic principles in robot team coordination and the definition of robotic decision support procedures, where the coordinating robot, endowed with a knowledge base, advises its peers. The new approaches will allow service robot developers to design coordination principles, where all robots select nondominated decisions taking into account the information about the coordinator's preference structure. The logical consistency of the decision making process is ensured when the individual preference structures of robots do not contradict that of the coordinator. The constructive algorithms, based on the above principles, have already been applied when designing decision engines for a team of autonomous harvesting robots [19]. The growing interest of industry and agriculture in the applications of AADSs and ANs in designing decision-making algorithms for mobile service robot teams is particularly promising.

References

1. Bolander, T., Dissing, L., Herrmann, N.: DEL-based Epistemic Planning for Human-Robot Collaboration: Theory and Implementation. In: Proceedings of the 18th International Conference on Principles of Knowledge Representation and Reasoning, Special Session on KR and Robotics, pp. 120–129 (2021). <https://doi.org/10.24963/kr.2021/12>
2. Buyukgoz, S., Grosinger, J., Chetouani, M., Saffiotti, A.: Two ways to make your robot proactive: reasoning about human intentions, or reasoning about possible futures, *Frontiers in Robotics and AI* 9 (2022). <https://doi.org/10.48550/arXiv.2205.05492>
3. Cao, X.-R., Zhang, J.: The n-th-order bias optimality for multichain Markov decision processes. *IEEE Trans. Autom. Control* 53(2), 496–508 (2008). <https://doi.org/10.1109/TAC.2007.915168>
4. Darwiche, A.: Modeling and reasoning with Bayesian networks. Cambridge University Press, p. 548 (2009). <https://doi.org/10.1017/CBO9780511811357>
5. Górecki, H., Skulimowski, A.M.J.: Safety principle in multiobjective decision support in the decision space defined by availability of resources. *Arch. Aut. i Telemech.* 11(2), 81–94 (1989)
6. Hughes, M.S., Lunday, B.J., Weir, J.D., Hopkinson, K.M.: The multiple shortest path problem with path deconfliction. *Eur. J. Oper. Res.* 292(3), 818–829 (2021). <https://doi.org/10.1016/j.ejor.2020.11.033>
7. Ji, T., Dong, R., Driggs-Campbell, K.: Traversing supervisor problem: an approximately optimal approach to multi-robot assistance. In: Proceedings of Robotics: Science and Systems, New York City, NY, USA (2022). <https://doi.org/10.15607/RSS.2022.XVIII.059>
8. Lauri, M., Hsu, D., Pajarinen, J.: Partially Observable Markov Decision Processes in Robotics: A Survey. *IEEE Trans. Rob.* 39(1), 21–40 (2023). <https://doi.org/10.1109/TRO.2022.3200138>
9. Liang, Z., Parlikad, A.K.: Predictive group maintenance for multi-system multi-component networks. *Reliability Eng. Syst. Saf.* 195, art. No. 106704, p.18 (2020). <https://doi.org/10.1016/j.res.2019.106704>
10. Rosen R.: *Anticipatory Systems: Philosophical, Mathematical and Methodological Foundations*, vol. 1. Pergamon Press, London, Ed.2. Springer (2012)
11. Skulimowski, A.M.J.: Solving vector optimization problems via multilevel analysis of foreseen consequences. *Found. Control Eng.* 10(1), 25–38 (1985)

12. Skulimowski, A.M.J.: Freedom of choice and creativity in multicriteria decision making. In: Theeramunkong, T., Kunifuji, S., Sornlertlamvanich, V., Nattee, C. (eds.) KICSS 2010. LNCS (LNAI), vol. 6746, pp. 190–203. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-24788-0_18
13. Skulimowski, A.M.J.: Anticipatory network models of multicriteria decision-making processes. *Int. J. Syst. Sci.* **45**(1), 39–59 (2014). <https://doi.org/10.1080/00207721.2012.670308>
14. Skulimowski, A.M.J.: Future prospects of human interaction with artificial autonomous systems. In: Bouchachia, A. (ed.) ICAIS 2014. LNCS (LNAI), vol. 8779, pp. 131–141. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-11298-5_14
15. Skulimowski, A.M.J.: Anticipatory control of vehicle swarms with virtual supervision. In: Hsu, Ch., Wang, S., Zhou, A., Shawkat, A. (eds.) IOV 2016: Nadi, Fiji, December 7–10, 2016, Proceedings. LNCS, vol. 10036, pp. 65–81. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-51969-2_6
16. Skulimowski, A.M.J.: Roadmapping collaborative exploitation and marketing of an AI-based knowledge platform. In: Reis, J.L., Parra López, E., Moutinho, L., Marques dos Santos, J.P. (eds.) Marketing and Smart Technologies, vol. 1. Smart Innovation, Systems and Technologies, vol. 279, Singapore, pp. 55–66. Springer (2022). https://doi.org/10.1007/978-981-16-9268-0_5
17. Skulimowski, A.M.J., Łydek, P.: Adaptive design of a cyber-physical system for industrial risk management decision support. In: Proceedings of the 17th IEEE International Conference on Control, Automation, Robotics and Vision (ICARCV), Dec. 11–13, 2022, Singapore, IEEE CPS, pp. 90–97 (2022). <https://doi.org/10.1109/ICARCV57592.2022.10004251>
18. Skulimowski, A.M.J., Badecka, I., Hassan, A., Kara, M., Łydek, P., Pukocz, P.: New methods of decision analysis and support and their applications in intelligent autonomous systems [in Polish], *Nauka-Technika-Technologia*, vol. 3, AGH Scientific Publishers, pp. 141–156 (2022). https://doi.org/10.7494/978-83-66727-83-0_9
19. Skulimowski, A.M.J., Pukocz, P., Badecka, I., Kara, M.: A novel software architecture of anticipatory harvesting robot teams. In: 2021 25th International Conference on Methods and Models in Automation & Robotics (MMAR), August 23–26, 2021, Międzyzdroje, Poland, IEEE, Piscataway, pp. 47–52 (2021). <https://doi.org/10.1109/MMAR49549.2021.9528474>
20. Soeffker, N., Ulmer, M.W., Mattfeld, D.C.: Stochastic dynamic vehicle routing in the light of prescriptive analytics: a review. *Eur. J. Oper. Res.* **298**(3), 801–820 (2022). <https://doi.org/10.1016/j.ejor.2021.07.014>
21. Weber, P., Medina-Oliva, G., Simon, C., Iung, B.: Overview on Bayesian networks applications for dependability, risk analysis and maintenance areas. *Eng. Appl. Artif. Intell.* **25**(4), 671–682 (2012). <https://doi.org/10.1016/j.engappai.2010.06.002>
22. Xidias, E.K., Azariadis, P.N.: Mission design for a group of autonomous guided vehicles. *Robot. Auton. Syst.* **59**(1), 34–43 (2011). <https://doi.org/10.1016/j.robot.2010.10.003>
23. Zhou, J., Coit, D.W., Felder, F.A., Wang, D.L.: Resiliency-based restoration optimization for dependent network systems against cascading failures. *Reliability Eng. Syst. Saf.* 207, art. No. 107383, p.18 (2021). <https://doi.org/10.1016/j.res.2020.107383>



Modeling of Thermal Processes in a Microcontroller System with the Use of Hybrid, Fractional Order Transfer Functions

Krzysztof Oprzędkiewicz^(✉), Maciej Rosół, and Wojciech Mitkowski

Department of Automatic Control and Robotics, Faculty of Electrical Engineering,
Automatic Control, Informatics and Biomedical Engineering,
AGH University of Science and Technology,
al. A Mickiewicza 30, 30-059 Krakow, Poland
kop@agh.edu.pl

Abstract. In the paper the problem of modeling of thermal processes in microcontroller system is addressed. The proposed models allow to describe thermal processes during work of an evaluation system. The temperature in critical places of the system is measured using thermal camera. The proposed models have the form of hybrid transfer functions, containing both Integer Order (IO) and Fractional Order (FO) parts. The step responses of the proposed transfer functions are computed analytically. Results of experiments show that the proposed models are more accurate in the sense of Mean Square Error (MSE) cost function than typical transfer function models with delay. The proposed models can be applied to predict of overheating of microcontroller systems.

Keywords: microcontroller · fractional order transfer function · thermal processes · thermal camera

1 Introduction

The fractional order calculus is a convenient tool to describe many complex physical phenomena. Non-integer models have been presented by many Authors, e.g. by [2–5, 11, 14]. Analysis of anomalous diffusion problem using fractional order approach and semigroup theory was presented for example by [12]. An observability problem for fractional order systems has been discussed e.g. by [7].

Fundamental information about heat processes in microcontroller systems is given e.g. in documentation [1]. This problem has been also considered in Ph. D. thesis [13]. Fractional order, constant parameter models describing a thermal behaviour of a microcontroller are also proposed by authors in conference presentation (MMAR 2022), but their accuracy is not fully satisfying.

The paper is organized as follows. Preliminaries give some elementary ideas from fractional calculus. Next the considered microcontroller system and its thermal behaviour are presented and the hybrid transfer function models are proposed. Finally proposed models are verified using experimental data.

2 Preliminaries

Elementary ideas from fractional calculus can be found in many books, for example: [3, 6, 10] or [11]. Here only some definitions necessary to explain of main results will be given.

Firstly the fractional-order, integro-differential operator is given (see for example [3, 8, 11]):

Definition 1 (*The elementary fractional order operator*). *The fractional-order integro-differential operator is defined as follows:*

$${}_a D_t^\alpha f(t) = \begin{cases} \frac{d^\alpha f(t)}{dt^\alpha} & \alpha > 0 \\ f(t) & \alpha = 0 \\ \int_a^t f(\tau)(d\tau)^\alpha & \alpha < 0. \end{cases} \tag{1}$$

where a and t denote time limits for operator calculation, $\alpha \in \mathbb{R}$ denotes the non integer order of the operation.

Next remember an idea of Gamma Euler function ([8]):

Definition 2. *The Gamma function*

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt. \tag{2}$$

Furthermore recall an idea of Mittag-Leffler functions. The two parameter Mittag-Leffler function is defined as follows:

Definition 3 (*The two parameter Mittag-Leffler function*)

$$E_{\alpha,\beta}(x) = \sum_{k=0}^\infty \frac{x^k}{\Gamma(k\alpha + \beta)}. \tag{3}$$

For $\beta = 1$ we obtain the one parameter Mittag-Leffler function:

Definition 4. *The one parameter Mittag-Leffler function*

$$E_\alpha(x) = \sum_{k=0}^\infty \frac{x^k}{\Gamma(k\alpha + 1)}. \tag{4}$$

The fractional-order, integro-differential operator (1) can be described by different definitions, given by Grünwald and Letnikov (GL Definition), Riemann and Liouville (RL Definition) and Caputo (C Definition). Only C definition will be employed in this paper. It is as follows [8]:

Definition 5. *The Caputo Definition of the FO operator*

$${}_0^C D_t^\alpha f(t) = \frac{1}{\Gamma(V - \alpha)} \int_0^\infty \frac{f^{(V)}(\tau)}{(t - \tau)^{\alpha+1-V}} d\tau. \tag{5}$$

In (5) V is an integer limiter of the non integer order: $V - 1 \leq \alpha < V \in \mathbb{N}$.

If $V = 1$ then consequently $0 \leq \alpha < 1$ is considered and the definition (5) takes the form:

Definition 6. *The Caputo definition of the FO operator for $0 \leq \alpha < 1$*

$${}_0^C D_t^\alpha f(t) = \frac{1}{\Gamma(1 - \alpha)} \int_0^\infty \frac{\dot{f}(\tau)}{(t - \tau)^\alpha} d\tau. \tag{6}$$

For the Caputo operator the Laplace transform can be defined [6]:

Definition 7. *The Laplace transform for Caputo operator*

$$\begin{aligned} \mathcal{L}({}_0^C D_t^\alpha f(t)) &= s^\alpha F(s), \quad \alpha < 0 \\ \mathcal{L}({}_0^C D_t^\alpha f(t)) &= s^\alpha F(s) - \sum_{k=0}^{v-1} s^{\alpha-k-1} {}_0 D_t^k f(0), \\ \alpha > 0, \quad v - 1 < \alpha \leq v \in \mathbb{N}. \end{aligned} \tag{7}$$

Finally the fractional order transfer function can be defined:

$$G(s) = \frac{b_m s^{\beta_m} + \dots + b_1 s^{\beta_1} + b_0}{a_n s^{\alpha_n} + \dots + a_1 s^{\alpha_1} + a_0}. \tag{8}$$

where α and β are fractional orders (commensurate or not commensurate), a and b are coefficients.

In modeling elementary forms of the general transfer function (8) are applied. The elementary FO transfer function used in this paper takes the following form:

$$G(s) = \frac{1}{T_\alpha s^\alpha + 1}. \tag{9}$$

The advantage of the elementary transfer function (9) is that the analytical formula of impulse and step responses can be given. The step response of this transfer function is as follows (see e.g. [2], p. 11):

$$y_{FO}(t) = \left(1(t) - E_\alpha \left(-\frac{t^\alpha}{T_\alpha} \right) \right). \tag{10}$$

In (10) $E_\alpha(\dots)$ is the one parameter Mittag-Leffler function (4).

3 The Considered Heat System

The tested development board STM32 Nucleo-F767ZI is shown in the Fig. 1. The Fig. 2 shows the system employed to supervise thermal processes during its work. The whole temperature field is registered with the use of thermal camera OPTRIS PI 450, equipped with dedicated lens O29. The camera is connected to computer via USB. The range of measurement is 0–250°C, the frame rate 80 Hz. The resolution of the camera’s sensor is 382 × 288 pixels. The camera is attached 300 mm over table, the applied lens gives field of view 165 × 121 mm with the size of the single pixel equal 0.43 × 0.42 mm. Data from camera are collected using dedicated software Optris PI Connect. The ambient temperature during experiments was equal 25 °C. It can be read at the border of the measuring area (see Fig. 4).

Methods of a numerical solving of partial differential equations using finite difference methods or finite element methods have been known and applied since years. However their application to modeling of heat processes in elements with more complex shapes results in a significant complication of a model, that is not needed, when we are interested only in modeling of the most critical, highest temperatures in selected points of a system.

The geometry of the considered evaluation board is illustrated by the Fig. 1. The model of the temperature of the whole plate and all attached elements employing classic numerical methods would be very complicated. Simultaneously from point of view of correct work of the system the most critical are temperatures of single elements. Such temperatures can be described using ordinary differential equations.

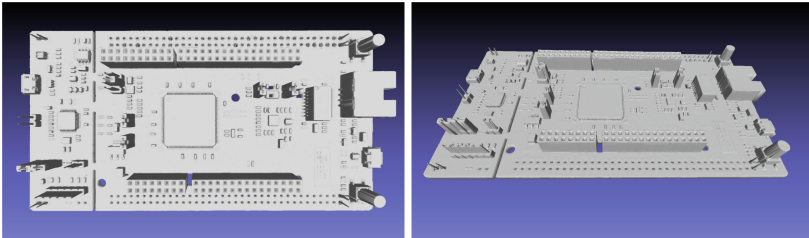


Fig. 1. The development board STM32.

During experiments the step response of the system was tested. The “zero” level is interpreted as the work of the STM32F767ZI system in the low power mode. This denotes the run of the empty while loop. The “one” level denotes the execution of calculations. The temperature we deal with is measured between start time $t_s > 0$ and final time $t_f > t_s$.

The temperature fields in the whole system for both states are illustrated by 3D diagrams shown in Figs. 3, 4 and contour plot given in Fig. 5. This plot shows also the points of temperature measurement.

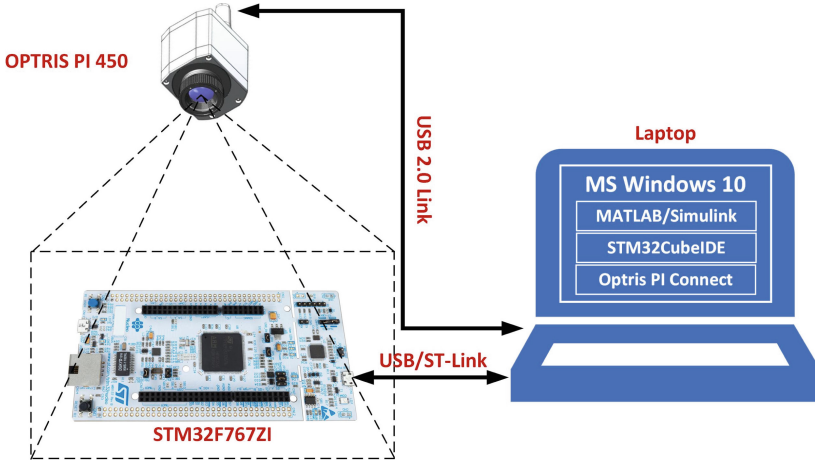


Fig. 2. The experimental system.

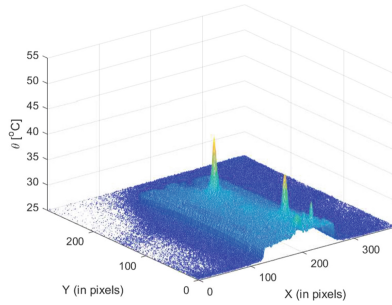


Fig. 3. The 3D temperature plot for “zero” level

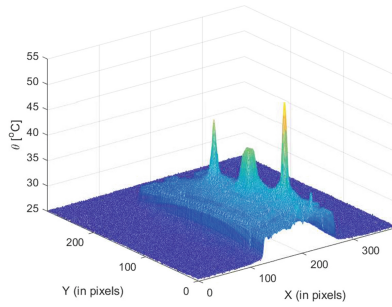


Fig. 4. The 3D temperature plot for “one” level.

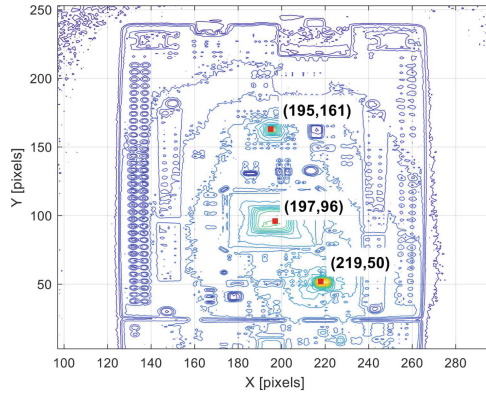


Fig. 5. The contour plot of temperature for “one” level and points of measurements.

For modeling were selected three points with maximum temperature pulses. Their coordinates x and y (in pixels) and detailed description of parts of the system are given in the Table 1. The time trends of temperature in these places are shown in Fig. 6.

Table 1. Coordinates of points of measurement (in pixels) and description of tested parts of the system.

point No	x	y	Description
1	219	50	Power Supply
2	197	96	Processor
3	195	161	Ethernet Port

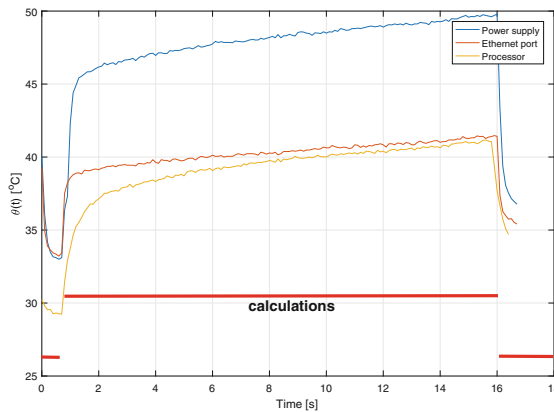


Fig. 6. The time trends of temperature in all tested places.

The shape of the step responses shown in Fig. 6 is somewhat surprising for the considered thermal process. The temperature rises rapidly at first, then slower but does not stabilize in a short time. The measurement was finished before reaching steady state because its continuation does not bring anything new and causes a significant increase in the amount of collected data.

After switching the processor to “zero” level, the temperature drops sharply.

Observed, unusual behavior of the above-described thermal process motivated us to propose new, hybrid transfer function models, containing both integer order (IO) and fractional order (FO) parts.

4 The Considered Transfer Function Models

4.1 Integer Order Transfer Functions with Delay

At the beginning the following, elementary, integer order (IO) transfer functions with delay were examined:

$$G_{IO1}(s) = \frac{k_1 e^{-\tau_1 s}}{T_1 s + 1}. \quad (11)$$

$$G_{IO2}(s) = \frac{k_2 e^{-\tau_2 s}}{(T_{21} s + 1)(T_{22} s + 1)}. \quad (12)$$

In (11) and (12) $k_{1,2}$ denote the steady-state gain, $\tau_{1,2}$ denote delay times and $T_{1,11,12}$, are time constants.

4.2 Fractional Order, Hybrid Transfer Functions

Next we propose the following, hybrid transfer functions, containing both IO and FO parts, analogically as in [9]. These functions are as follows:

$$G_{FO1}(s) = \frac{k_1}{(T_1 s + 1)(T_{\alpha_1} s^{\alpha_1} + 1)}. \quad (13)$$

$$G_{FO2}(s) = \frac{k_2}{(T_{11} s + 1)(T_{12} s + 1)(T_{\alpha_2} s^{\alpha_2} + 1)}. \quad (14)$$

5 Experimental Validation of Models

The quality of all considered models was examined using known MSE cost function:

$$MSE = \frac{1}{K} \sum_{k=1}^K (y_{IO,FO}(kh) - y_e(kh))^2. \quad (15)$$

In (15) K is the number of all collected samples, h is the sample time, $y_{IO,FO}(kh)$ and $y_e(kh)$ denote step responses of model and real system respectively.

The heating of the system was tested during evaluation of FOPID control algorithm with the use of PSE approximation. The source code of this testing program is shown in the Fig. 7.

```

for( uint16_t j=0; j <= L_PSE; j++)
{
    y_PSE_Der[L_PSE] = y_PSE_Der[L_PSE] + PSE_Coeff_Der[j]*u_PSE[L_PSE-j];
    y_PSE_Int[L_PSE] = y_PSE_Int[L_PSE] + PSE_Coeff_Int[j]*u_PSE[L_PSE-j];
}
|
Yout_Der = Td*( 1.0/pow(T0,FracOrder_Der) )*y_PSE_Der[L_PSE];
Yout_Int = Ti*( 1.0/pow(T0,FracOrder_Int) )*y_PSE_Int[L_PSE];
Y_PID = Kp*u_PSE[L_PSE] + Yout_Int + Yout_Der;

memcpy ( y_PSE_Der, y_PSE_Der+1, L_PSE*sizeof(double));
memcpy ( y_PSE_Int, y_PSE_Int+1, L_PSE*sizeof(double));
memcpy ( u_PSE, u_PSE+1, L_PSE*sizeof(double));
y_PSE_Der[L_PSE] = 0;
y_PSE_Int[L_PSE] = 0;
    
```

Fig. 7. Source code used during experiment.

5.1 Integer Order Models with Delay

Firstly the IO transfer functions with delay (11) and (12) were examined. Delay was modeled using MATLAB function *pade*, step responses of models were computed using MATLAB function *step*. Results are described in Tables 2, 3 and illustrated by Figs. 8 and 9.

Table 2. Results of identification of the transfer function (11).

Element	τ [s]	T [s]	MSE
Power supply	0.5521	5.8423	1.0448
Processor	2.9987	12.8423	0.6044
Eth. port	0.2999	4.8722	0.4353

Table 3. Results of identification of the transfer function (12).

Element	τ_2 [s]	T_1 [s]	T_2 [s]	MSE
Power supply	0.1240	1.5515e-07	6.5106	1.0246
Processor	2.4551	0.1987	14.9834	0.6409
Eth. port	0.3651	0.1951	2.7334	0.4743

5.2 Hybrid FO-IO Models

Next the proposed hybrid models (13) and (14) were tested. The step responses of both transfer functions were computed using analytical formula (10) and MATLAB function *lsim* used to IO part. The method of calculating was following.

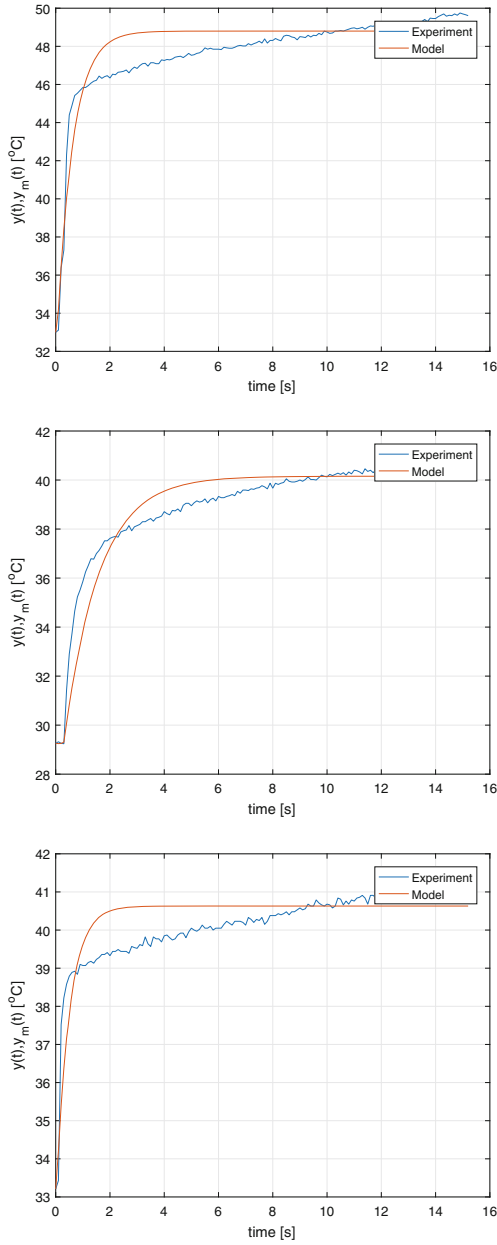


Fig. 8. Step responses of the 1'st order model with delay (11) vs experiments. From top to bottom: power supply, processor, ethernet port.

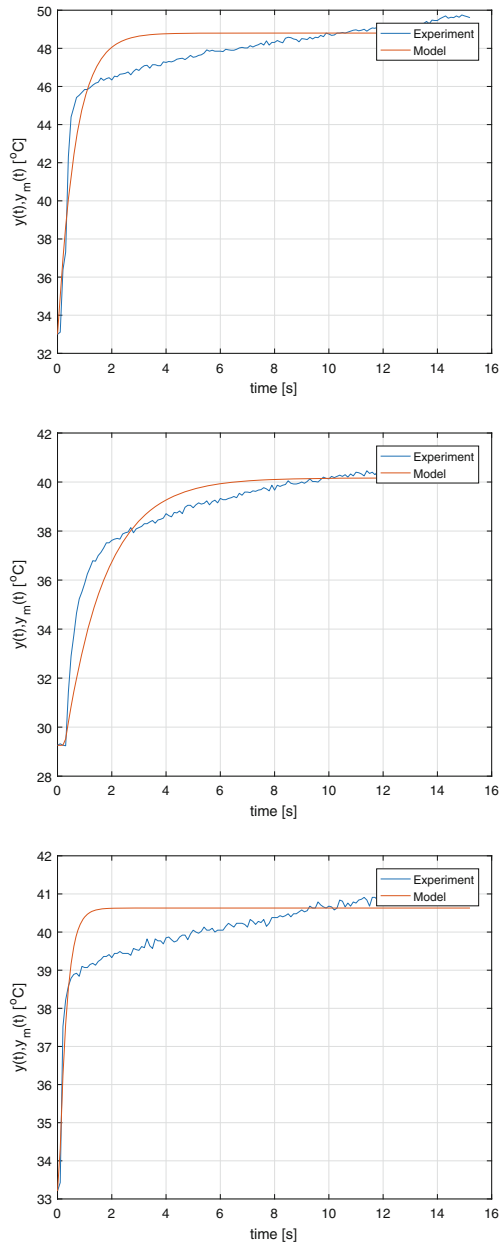


Fig. 9. Step responses of the 2nd order model with delay (12) vs experiments (blue). From top to bottom: power supply, processor, ethernet port.

Firstly the step response of FO part was calculated using formula (10) and at the same time grid as used in experiments. Next the output signal from FO part is employed as the input signal to compute response of IO part using MATLAB function *lsim*. Such an approach assures the good accuracy of computations in contrast to use approximations (e.g. ORA).

Results of identification are presented in Tables 4, 5 and illustrated by Figs. 10, 11.

Table 4. Results of identification of the transfer function (13).

Element	α_1	T_{α_1} [s]	T [s]	MSE
Power supply	0.1497	19.0251	0.2501	0.0945
Processor	0.4625	2.7154	0.2718	0.2377
Eth. port	0.2125	20.0021	0.0524	0.0112

Table 5. Results of identification of the transfer function (14).

Element	α_2	T_{α_2} [s]	T_{11} [s]	T_{12} [s]	MSE
Power supply	0.1406	22.2249	0.1213	0.1214	0.0438
Processor	0.3709	2.9037	0.2645	0.2645	0.0213
Eth. port	0.2002	22.5031	2.59e-09	0.0674	0.0094

6 Discussion of Results and Final Conclusions

The summary of the results will be started by the comparing of values of the cost function (15) for all tested models and elements of the system. This is presented in the Table 6.

Table 6. Cost function (15) for all tested models and elements.

Element, model	IO model(11)	IO model (12)	IO-FO model (13)	IO-FO model (14)
Power supply	1.0448	1.0246	0.0945	0.0438
Processor	0.6044	0.6409	0.2377	0.0213
Eth. port	0.4353	0.4743	0.0112	0.0094

The Table 6 shows that the proposed, hybrid transfer function models more accurately describe the thermal processes in the microcontroller system than the

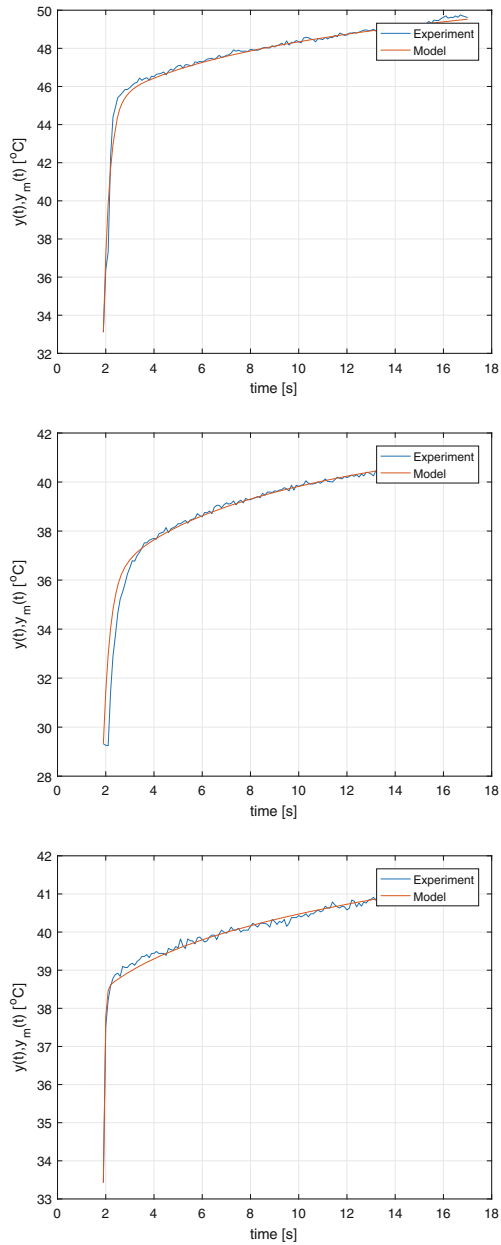


Fig. 10. Step responses of the hybrid model (13) vs experiments. From top to bottom: power supply, processor, ethernet port.

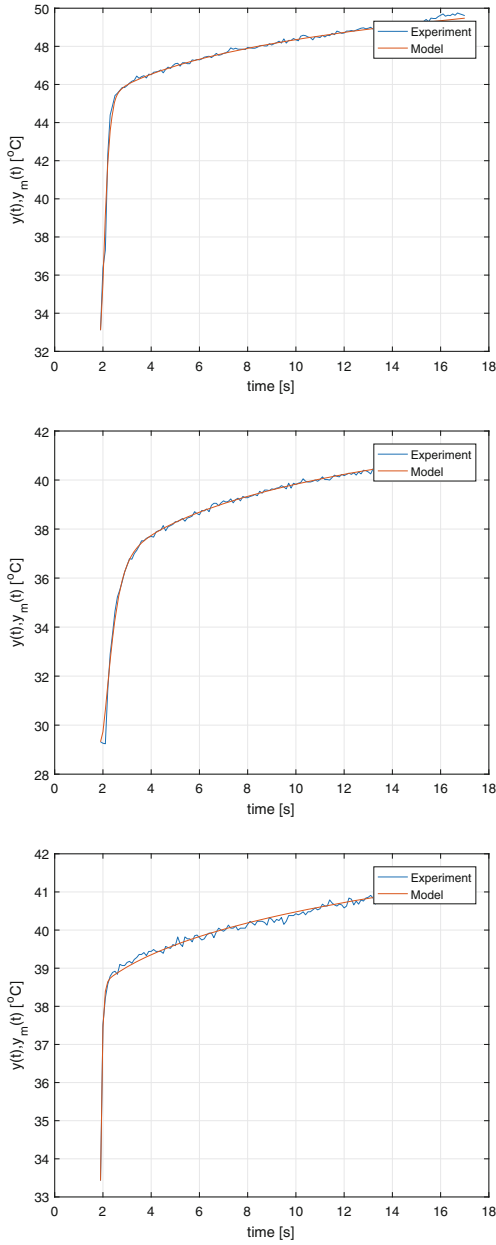


Fig. 11. Step responses of the hybrid model (14) vs experiments. From top to bottom: power supply, processor, ethernet port.

classic IO models with delay. This can be explained by the fact that the Mittag-Leffler function (4) better describes the behaviour of the temperature of tested elements of the system.

Next, the hybrid model using 2nd order IO part is more accurate than the model using 1st order IO part.

The proposed models are relatively easy in use and accurate due to calculate of the step responses using analytical formulae.

The spectrum of further investigations of the presented issue is broad.

Firstly the varying emissivity of different part of tested system generates problems during thermal imaging. In the proposed model it can be described e.g. by the use of interval parameters: time constant T and fractional order α . Such a model is currently being developed.

Next, from point of view of practice is to test the proposed models for different types of calculations, CPU loads, use of memory and so on.

Another issue is the building of a fractional order state space model as well as development and testing of discrete-time models, easy to digital implementation. Such a discrete model can be applied e.g. to implementation of a self-diagnostic system to warn of overheating.

From point of view understanding of phenomena taking place in the tested system interesting is also a measurement of the electrical power supplied to the board and associating with thermal processes.

Acknowledgments. This paper was sponsored by AGH UST project no 16.16.120.773.

References

1. Thermal management guidelines for STM32 applications (2019)
2. Caponetto, R., Dongola, G., Fortuna, L., Petras, I.: Fractional order systems: modeling and control applications. In: Chua, L.O. (ed.) World Scientific Series on Non-linear Science, pp. 1–178. University of California, Berkeley (2010)
3. Das, S.: Functional Fractional Calculus for System Identification and Controls. Springer, Heidelberg (2010). <https://doi.org/10.1007/978-3-540-72703-3>
4. Dzieliński, A., Sierociuk, D., Sarwas, G.: Some applications of fractional order calculus. Bull. Polish Acad. Sci. Tech. Sci. **58**(4), 583–592 (2010)
5. Gal, C.G., Warma, M.: Elliptic and parabolic equations with fractional diffusion and dynamic boundary conditions. Evol. Eqns. Control Theory **5**(1), 61–103 (2016)
6. Kaczorek, T.: Selected Problems of Fractional Systems Theory. LNCIS, Springer, Heidelberg (2011). <https://doi.org/10.1007/978-3-642-20502-6>
7. Kaczorek, T.: Reduced-order fractional descriptor observers for a class of fractional descriptor continuous-time nonlinear systems. Int. J. Appl. Math. Comput. Sci. **26**(2), 277–283 (2016)
8. Kaczorek, T., Rogowski, K.: Fractional Linear Systems and Electrical Circuits. Bialystok University of Technology, Bialystok (2014)
9. Oprzędkiewicz, K., Gawin, E., Mitkowski, W.: Application of fractional order transfer functions to modeling of high - order systems. In: MMAR, 20th international conference on Methods and Models in Automation and Robotics: 24–27 August 2015. Międzyzdroje, Poland, pp. 1169–1174 (2015)

10. Ostalczyk, P.: Discrete Fractional Calculus. Applications in Control and Image Processing. World Scientific, New Jersey, London, Singapore (2016)
11. Podlubny, I.: Fractional Differential Equations. Academic Press, San Diego (1999)
12. Popescu, E.: On the fractional Cauchy problem associated with a feller semigroup. *Math. Rep.* **12**(2), 181–188 (2010)
13. Sankaranarayanan, K.: Thermal modeling and management of microprocessors. Ph.D. thesis, School of Engineering and Applied Science University of Virginia, Virginia, USA (2009)
14. Sierociuk, D., et al.: Diffusion process modeling by using fractional-order models. *Appl. Math. Comput.* **257**(1), 2–11 (2015)



Identification of Magnetorheological Damper Model for Off-Road Vehicle Suspension

Piotr Krauze¹(✉), Marek Płaczek², Zbigniew Żmudka³, Dawid Bauke¹, Przemysław Olszówka², Jakub Turek², Artur Wyciślok¹, Maciej Ziaja¹, Szymon Zosgórnik¹, Wojciech Janusz⁵, Grzegorz Przybyła³, and Michał Wychowański⁴

¹ Silesian University of Technology, Faculty of Automatic Control, Electronics and Computer Science, Akademicka 16, 44-100 Gliwice, Poland

piotr.krauze@polsl.pl

² Silesian University of Technology, Faculty of Mechanical Engineering, Konarskiego 18A, 44-100 Gliwice, Poland

marek.placzek@polsl.pl

³ Silesian University of Technology, Faculty of Energy and Environmental Engineering, Konarskiego 18, 44-100 Gliwice, Poland

zbigniew.zmudka@polsl.pl

⁴ Engineering Office, 27-530 Ożarów, Poland

wychowanski@hbm.com.pl

⁵ Gliwice, Poland

Abstract. The article presents results of experimental identification of MR (magnetorheological) damper model for an off-road vehicle. Original shock-absorbers of the vehicle were replaced by suspension MR dampers supplied by a dedicated measurement and control system. The vehicle includes several sensors which allow to observe electric currents controlling MR dampers, forces generated in the front vehicle suspension columns and other sensors which measure deflection of these suspension parts. The front part of the semi-active suspension was tested using laboratory setup with vibration mechanical exciters generating sinusoidally variable excitation at frequency equal to 12 Hz. MR dampers were directly controlled by peripheral electronic units generating different constant currents equal to 0, 0.11 and 0.37 A. As a result, evaluated force-velocity characteristics and analysis of MR damper behaviour in the target vehicle were carried out. Further steps of analysis included mathematical description of the selected MR damper using several types of MR damper models, i.e.: Bingham and hyperbolic tangent function models as well as their variants including first-order output dynamics. It was shown that the dynamic model based on hyperbolic tangent function offers the best fitness to the response of the actual MR damper with respect to the mean-squared error.

1 Introduction

Hazardous and varying road conditions promote adaptive designs of vehicle suspension over non-adaptive ones. An MR (magnetorheological) damper [1] is an example of semi-active damper apart from other types such as servo-valve [2] or electrorheological [3]. Comparing to classical shock absorber the MR damper is filled with MR fluid which consists of magnetizable particles suspended in oil [4]. The MR fluid flows through damper piston gaps where it is subjected to magnetic field induced by electric coils built into the piston. MR fluid particles reorganize under magnetic field in chain-like structures which counteract the fluid flow and increase damping parameter in macroscopic scale. The MR damper allows to control the energy dissipation in vehicle suspension and it is favoured for low energy consumption and fast response time in comparison to active suspension with force generators [5].

Possible control methods applied for MR dampers commonly consist of several control layers dedicated to electric current supplying MR damper, generation of MR damper force or mitigation of vibration of selected vehicle parts. Two approaches to force control layer can be generally distinguished, i.e. open loop [6] control algorithms based on inverse MR damper model or closed loop methods based on additional force measurements [7,8]. Both of above presented approaches have their disadvantages. On the one hand the force sensor needs to be carefully integrated into the suspension column in order not to weaken suspension design and maintain sufficient driving safety. On the other hand the open loop control approach is commonly based on indirect estimation of damper force achieved based on a previously identified MR damper model. Thus, the latter algorithm can be influenced by the fact that vehicle or suspension itself can operate not always in nominal conditions but also in conditions which are not covered by the applied damper model.

Behaviour of MR damper is mostly analysed based on characteristics of damper force presented with respect to piston velocity. Such characteristics evaluated for different operating conditions reveal force saturation for higher piston velocities and hysteresis loops [9]. These phenomena make modelling difficult and, thus, numerous approaches to modelling of MR damper behaviour are still being presented in the literature. The classical Bingham model [10] consists of a Coulomb force which mainly maps nonlinear behaviour and force saturation of MR damper. An additional viscous damping component describes force generated for higher piston velocities. Other possible heuristic methods try to more accurately describe nonlinear shape of force-velocity characteristics, e.g. based on hyperbolic tangent function [9] or polynomial models [11]. Additionally, hysteresis loop is studied in the literature and made dependent on piston acceleration [12] or described using the Bouc-Wen model [13]. Other possible approaches to modelling of hysteresis loop assume that this behaviour is caused by dynamic response of MR damper and thus can be modelled using first-order output dynamics [14] or it can be described by desired characteristics of hypothetical all-pass filters [15].

Common studies presented in the literature related to modelling of MR damper behaviour are based on simulation results or measurements taken for MR damper which is separately tested using MTS (material testing system). Thus, the study discussed in the current article related to operation of the MR damper in an experimental off-road vehicle can be treated as one of the main contribution. It gives an insight into the behaviour of MR dampers applied in the target automotive application. Furthermore, the article presents MR damper model identification carried out assuming that the hysteresis behaviour is described by the output first-order dynamics. Such modelling approach could be more directly integrated in future developments of semi-active control algorithms. The article consists of five sections where the Sect. 1 reports state-of-the-art in the field of modelling and identification of MR damper behaviour. In the Sect. 2 the experimental vehicle is described as well as the measurement and control system installed in the vehicle is discussed. The Sect. 3 defines different approaches and models of MR dampers applied in further identification process. In the Sect. 4 results of identification of selected MR damper models are presented and finally remarks studied in the article are concluded in Sect. 5.

2 Measurement and Control System for the Experimental Off-Road Vehicle

The experimental off-road vehicle is a key element of the presented experimental setup (Fig. 1). Characteristic feature of the vehicle is an independent suspension of its each wheel. Four original shock absorbers of this vehicle were replaced by suspension MR dampers manufactured by Lord Corporation. The vehicle is equipped with a dedicated measurement and control system responsible for



Fig. 1. Experimental off-road vehicle and the diagnostics station with mechanical vibration exciters

tracking motion of selected vehicle parts and for generation of control signals desired for the installed MR dampers.

The measurement part of the system consists of numerous sensors where the following measurements are used within the presented study:

- electric control current supplying suspension MR dampers generated as PWM (pulse width modulation) voltage signal measured by a dedicated inductive current sensors,
- force generated in each of the front suspension columns measured axially by force sensors which are installed into these suspension columns using dedicated mechanical adapters,
- suspension deflection of each of the front suspension columns measured axially by LVDT (linear variable differential transformer) displacement sensors which are similarly installed using dedicated adapters.

Flow of measurement and control data consists of several nodes. During experiments an operator supervises a main suspension controller and an additional measurement controller. The main suspension controller transfers MR damper control signals to selected peripheral control units which are responsible for generation of physical control currents. The additional measurement device is directly connected to sensors listed above which simultaneously track operation of MR damper and motion of the vehicle suspension.

Presented experimental results were obtained for the vehicle subjected periodic road excitation generated by simultaneously operating right and left mechanical exciters. These exciters are separately driven by electric motors controlled with inverters. Operation of inverters is additionally supervised and synchronized with a dedicated microcontroller device. Procedure of operation of mechanical exciters during each experiment can be distinguished into several

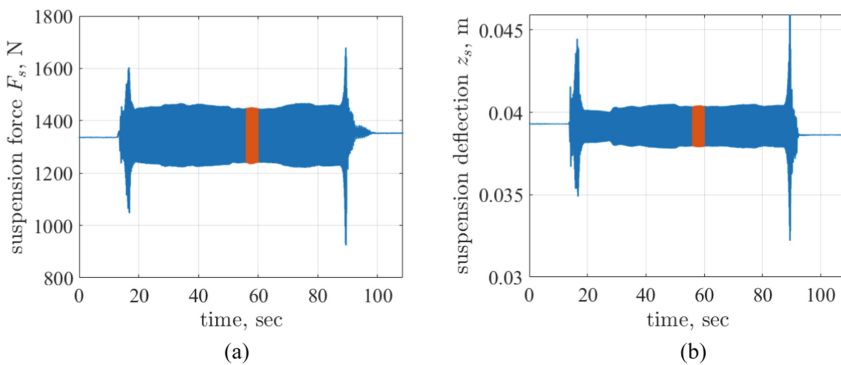


Fig. 2. Time diagrams presenting procedure of experiments carried out using the diagnostics station: a) force generated in the front left suspension column, b) deflection of the front left suspension part. Further analysis is based on parts of measurement data marked in red.

phases They can be indicated based on time diagrams presenting force generated in the front left suspension part and its deflection in Fig. 2.

Initially, mechanical exciters are activated and excitation frequency increases 15 Hz passing through their resonance frequency approximately equal to 8.5 Hz. The resonance is indicated by a significant increase of amplitudes of both suspension deflection and damper force signals. Mechanical exciters stabilize their operation at this frequency during 15 s and then frequency decreases to the 12 Hz. During this second phase, which lasts 60 s, operation of suspension MR dampers is analysed. Simultaneously, the upper layer controller of mechanical exciters tries to equalize phases between vibration signals generated by both exciters as possible. Finally, the mechanical exciters are deactivated and prepared for the next experiment while their heating is examined using separate temperature sensors.

Several observations can be made based on Fig. 2. The static deflection of the suspension is equal to 39 mm while static force generated in the suspension column is equal to 1334 N for such conditions. It is worth noting that the corresponding static suspension force is much greater than the weight on the front left wheel of the vehicle (approximately equal to 100 kg). It is caused by the fact that the suspension column is significantly inclined relative to the vertical and such force measured in the suspension additionally includes a horizontal force component which pushes the wheels of the vehicle sideways.

During selected experiment (Fig. 2) amplitude of suspension deflection varied from 1 mm for the target frequency 12 Hz to 7 mm for the resonance occurrence during deactivation of mechanical exciters. In the case of force measurements amplitude stabilized at approximately 120 N for the target frequency while it can reach approximately 380 N during the end of experiment. The whole single experiment lasted approximately 110 s.

3 Modelling of Magnetorheological Damper Behaviour

Further analysis of the front left suspension MR damper is carried out for 5 s long parts of measurement data as marked in red in Fig. 2. The selected measurement signals were processed in the following phases:

- acquisition of measurement data with sampling frequency equal to 10 kHz in order to avoid aliasing effect,
- signal processing using higher-order lowpass digital filter 100 Hz cut-off frequency in order to exclude higher frequency disturbances,
- signal decimation in order to decrease the sampling frequency down to 1 kHz,
- calibration of measurements based on sensors' parameters and subtraction of the constant component,
- estimation of damper piston relative velocity.

The damper piston velocities denoted as v_{mr} and assumed due to suspension construction as equivalent to suspension velocities, denoted as v_s , were estimated

based on suspension deflection measurements denoted as z_s using a straightforward second-order digital filter:

$$v_{\text{mr}} = v_s = \frac{f_s}{2}(1 - z^{-2}) \cdot z_s, \quad (1)$$

where z^{-1} denotes the delay operator and f_s denotes the sampling frequency. The delay of the filter was finally compensated since the presented FIR (finite impulse response) filter exhibits a constant, single sample group delay for the whole frequency range.

3.1 Analysis of the Actual MR Damper Behaviour

As a result of above-mentioned phases of signal processing a comparison of force-velocity characteristics for the front left suspension column can be presented in Fig. 3. The force-velocity characteristics were evaluated for different control currents, i.e. 0, 0.11 or 0.37 A supplying both front suspension MR dampers. Due to the construction of the suspension column, the resultant force consists of force component generated by the suspension spring $-k_s \cdot z_s$ and the MR damper $F_{\text{mr,model}}$, as follows:

$$F_{s,\text{model}} = F_{\text{mr,model}} - k_s \cdot z_{\text{mr}}. \quad (2)$$

Parameter of the suspension spring was independently measured and is equal to $k_s=22208 \text{ Nm}^{-1}$. However, it can be indicated based on presented characteristics that force generated by MR damper is the dominant one.

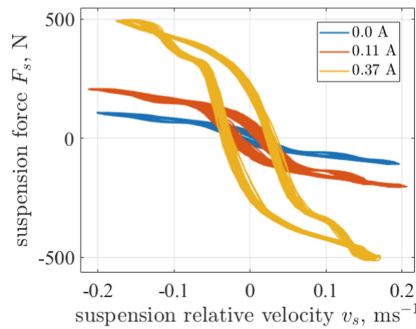


Fig. 3. Comparison of force-velocity characteristics obtained from measurements of the front left suspension column for different control currents

Force-velocity characteristics show nonlinear behaviour of MR damper in the form of force saturation visible for higher velocities, and more clearly for lower control currents. Hysteresis loop is more clearly revealed for higher control currents. Maximum forces generated by the suspension column vary from 106 to 500 N while the amplitude of damper piston relative velocity varies from 0.17 to 0.21 ms^{-1} .

3.2 Selected MR Damper Models

Four variants of MR damper modelling approaches were analysed in the presented article. Force generated according to first selected MR damper model, i.e. the Bingham model which consists of a Coulomb force and viscous damping component, denoted as $F_{\text{mr,bh}}$ can be defined as follows:

$$F_{\text{mr,bh}} = -f_{\text{bh}} \cdot \text{sign}(v_{\text{mr}}) - c_{\text{bh}} \cdot v_{\text{mr}}, \quad (3)$$

where Coulomb force and viscous damping parameters of the model are denoted as f_{bh} and c_{bh} , respectively.

The second considered MR damper model is based on hyperbolic tangent function and its generated force $F_{\text{mr,th}}$ can be defined as follows:

$$F_{\text{mr,th}} = -\alpha_{\text{th}} \cdot \tanh(\beta_{\text{th}} \cdot v_{\text{mr}}) - c_{\text{th}} \cdot v_{\text{mr}}, \quad (4)$$

where parameters of the model are denoted as α_{th} , β_{th} and c_{th} .

Each of the above presented MR damper models was also tested assuming existence of an additional output dynamics defined by the following transfer function:

$$K_{\text{mr}} = \frac{1}{1 + sT_{\text{mr}}}, \quad (5)$$

where s is an operator of the Laplace transform and T_{mr} denotes the time constant of the output dynamics of MR damper model.

4 Results of Identification of Selected MR Damper Models

Identification of the presented suspension models was carried out using *fmincon* function implemented in Matlab environment. Goal of the identification was defined as minimization of the following cost function:

$$J = \sqrt{\frac{\sum^N (F_{\text{mr}} - F_{\text{model}})^2}{\sum^N F_{\text{mr}}^2}}, \quad (6)$$

which is mainly the mean-squared error evaluated between measurements denoted as F_{mr} and response of the model F_{model} for a given set of parameters and number of samples N . The additional components of the square root and the normalization by mean-squared value of measured force do not influence the identification process and were applied only for further analysis of identification results.

Results of identification for different classes of MR damper models are compared in Table 1 based on values of cost function J . Quality of fitness of model responses to measurement data is inversely proportional to the value of cost function J . It can be noticed that best results were obtained for MR damper

Table 1. Comparison of values of identification cost function J evaluated for different control currents i_{mr} and selected MR damper models. The best identification results are bolded.

i_{mr}	Bingham model	\tanh model	dynamic Bingham model	dynamic \tanh model
0.0 A	0.189	0.176	0.190	0.176
0.11 A	0.142	0.100	0.131	0.100
0.37 A	0.202	0.174	0.097	0.077

dynamic model - MR damper model with first-order output dynamics defined by Eq. 5

models based on hyperbolic tangent function. Furthermore, application of output dynamics is recommended for higher control currents.

Comparison of force-velocity characteristics obtained from measurements and responses of \tanh MR damper model including output dynamics are presented in Fig. 4. It can be indicated based on force-characteristics and values of J presented

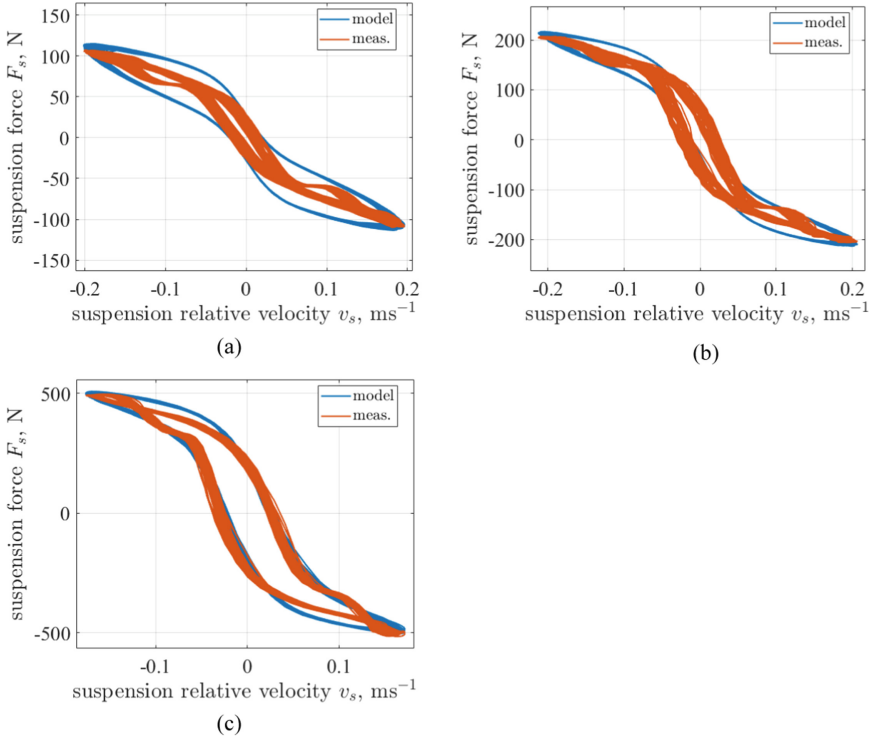


Fig. 4. Comparison of force-velocity characteristics obtained from measurements of the front left suspension column and from responses of \tanh MR damper model including output dynamics, evaluated for different control currents: a) 0.0 A, b) 0.11 A, c) 0.37 A.

in Table 1 that for higher control currents MR damper model is better fitted to measurement data. Stiffness parameters of the suspension spring k_s was taken into account for all identification results. It is clearly visible for the lowest control current where response of MR damper model exhibits a kind of an oval shape for extreme values of the suspension relative velocities.

5 Conclusions

The article presents results of experimental identification of MR (magnetorheological) damper model carried out in the off-road vehicle. The front part of the semi-active suspension was tested using laboratory setup equipped with vibration mechanical exciters generating sinusoidally variable excitation at frequency equal 12 Hz. MR dampers were controlled by peripheral electronic units which directly generate different constant currents equal to 0, 0.11 and 0.37 A. As a result, evaluated force-velocity characteristics and analysis of MR damper behaviour in the target vehicle application were carried out based on measured deflection and force generated in the front left suspension column.

The study of measurement results of MR damper taken in the experimental off-road vehicle can be treated as one of the main contribution which give insight into the behaviour of MR dampers applied in the target automotive application. Further steps of the analysis included mathematical description of MR damper behaviour using several types of MR damper models, i.e.: Bingham and hyperbolic tangent function models as well as their variants including first-order output dynamics.

It was shown that the MR damper model based on hyperbolic tangent function offers the best fitness to the response of the actual MR damper with respect to the mean-squared error quality index. Furthermore, application of the output dynamics is recommended for higher control currents which gave better fitness of the MR damper model to measurement data. Future research will be focused on application and testing of the presented identification approach for different driving conditions as well as it can be integrated with suspension control algorithms.

Acknowledgements. The research reported in this paper is the result of the PBL project co-financed by the European Union from the European Social Fund in the framework of the project “Silesian University of Technology as a Center of Modern Education based on research and innovation” POWR.03.05.00-00-Z098/17.

References

1. Spencer, B.F., Dyke, S.J., Sain, M.K., Carlson, J.D.: Phenomenological model of a magnetorheological damper. *ASCE J. Eng. Mech.* **123**, 230–238 (1997)
2. Soliman, A.M.A., Kaldas, M.M.S.: Semi-active suspension systems from research to mass-market - a review. *J. Low Frequency Noise Vibr. Active Control* **40**(2), 1005–1023 (2019)

3. Symans, M.D., Constantinou, M.C.: Semi-active control systems for seismic protection of structures: a state-of-the-art review. *Eng. Struct.* **21**, 469–487 (1999)
4. Sapiński, B.: *Magnetorheological Dampers in Vibration Control*. AGH University of Science and Technology Press, Cracow (2006)
5. Koo, J.-H., Goncalves, F.D., Ahmadian, M.: A comprehensive analysis of the response time of MR dampers. *Smart Mater. Struct.* **15**, 351–358 (2006)
6. Krauze, P., Kasprzyk, J.: Driving safety improved with control of magnetorheological dampers in vehicle suspension. *Appl. Sci.* **10**(24), 8892, 1–29 (2020)
7. Terasawa, T., Sano, A.: Fully adaptive vibration control for uncertain structure installed with MR damper. In: *Proceedings of American Control Conference 2005*. Portland, OR, USA, 8–10 June 2005
8. Mori, T., Nilkhamhang, I., Sano, A.: Adaptive semi-active control of suspension system with MR damper. In: *Proceedings of 9th IFAC Workshop on Adaptation and Learning in Control and Signal Processing 2007*, vol. 9, pp. 191–196 (2007)
9. Kasprzyk, J., Wyrwał, J., Krauze P.: Automotive MR damper modeling for semi-active vibration control. In: *Proceedings of International Conference on Advanced Intelligent Mechatronics*. Besancon, France, 8–11 July 2014
10. Sapiński, B.: Parametric identification of MR linear automotive size damper. *J. Theor. Appl. Mech.* **40**(3), 703–722 (2002)
11. Dong, X.M., Yu, M., Li, Z., Liao, C., Chen, W.: A comparison of suitable control methods for full vehicle with four MR dampers, part I: formulation of control schemes and numerical simulation. *J. Intell. Mater. Syst. Struct.* **20**, 771–786 (2009)
12. Kasprzyk, J., Plaza, K., Wyrwał, J.: Identification of a magnetorheological damper for semi-active vibration control. In: *Proceedings of International Congress on Sound and Vibration*. Vilnius, Lithuania, 8–12 July 2012
13. Ogonowski, S., Krauze, P.: Trajectory control for vibrating screen with magnetorheological dampers. *Sensors* **22**(11), 4225, 1–33 (2022)
14. Song, X., Ahmadian, M., Southward, S.C.: Modeling magnetorheological dampers with application of nonparametric approach. *J. Intell. Mater. Struct.* **16**, 421–432 (2005)
15. Krauze, P., Wyrwał, J.: Magnetorheological damper dedicated modelling of force-velocity hysteresis using all-pass delay filters. In: *Swiątek, J., Grzech, A., Swiątek, P., Tomczak, J.M. (eds.) Advances in Systems Science*. AISC, vol. 240, pp. 425–433. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-01857-7_41



Physics-Informed Hybrid Neural Network Model for MPC: A Fuzzy Approach

Krzysztof Zarzycki^(✉) and Maciej Ławryńczuk

Institute of Control and Computation Engineering, Warsaw University of Technology,
ul. Nowowiejska 15/19, 00-665 Warsaw, Poland
{Krzysztof.Zarzycki,Maciej.Lawrynczuk}@pw.edu.pl

Abstract. In practice, data from the process typically can be collected only for limited ranges of the entire area of process operation, which limits the quality of data-driven models. On the other hand, a fundamental process model may be available but flawed due to measurement noise or immeasurable model parameters. We recommend the described Physics-Informed Hybrid Neural Network (PIHNN) model to model the process precisely. It is comprised of a series of black-box sub-models; we use Gated Recurrent Unit networks (GRU) in this role. Neural and fundamental sub-models are combined using a fuzzy data fusion block. We report the model development procedure for a benchmark pH reactor and show that the model gives very good accuracy. Finally, the PIHNN model is implemented in a Model Predictive Control algorithm.

Keywords: Model Predictive Control · GRU Neural Networks · Physics-Informed Neural Networks

1 Introduction

Model Predictive Control (MPC) [5, 14] is an advanced control technique used when classical methods fail. Among the most significant advantages of MPC, we stress good control quality for multivariable and nonlinear processes and the ability to respect different types of constraints. In practice, MPC finds applications in embedded systems, e.g., [1, 11], and in industry, e.g., [3, 7].

A precise model is crucial for MPC. There are two main modelling approaches. Firstly, First Principle (FP) models rely on physical laws and are composed of differential equations. They may offer excellent modelling quality, but a thorough knowledge of the phenomena taking place and the values of parameters is required. Secondly, data-driven models are determined entirely using measurements of process variables. Examples of black-box models are Support Vector Machines (SVM) [9], and neural networks, e.g., Multi-Layer Perceptrons (MLP) [8], Radial Basis Function (RBF) networks [4], Long Term Short Memory (LSTM) [13], and Gated Recurrent Unit networks (GRU) [16]. Identification of black-box models does not require fundamental knowledge, but in practice, data sets for some neighbourhoods of operating points are only available.

Physics-Informed Neural Networks (PINNs) combine physical laws with data-driven methods. PINNs may be used for processes described by Ordinary Differential Equations (ODEs), where some model parameters are immeasurable [12], or subject to inaccuracies [2], and systems governed by Partial Differential Equations (PDEs) [15].

This work deals with a modelling task fraught with problems often encountered in practice. Firstly, we assume that an imperfect FP model is available. Secondly, process measurements are available from only a few narrow ranges of the process operation. As a result, data-driven models only work correctly locally. This paper proposes a new Physics-Informed Hybrid Neural Network (PIHNN) framework based on GRU neural networks. It combines the advantages of data-driven and FP modelling approaches using fuzzy data fusion. As a result, the proposed PIHNN has good modelling quality for the full range of process operation. For a benchmark pH reactor, we report a complete model development procedure. In particular, we study the impact of different fuzzy membership functions and tuning methods on model accuracy. We compare the results of fuzzy data fusion with a simple average model. Finally, the selected PIHNN model is implemented in the MPC algorithm.

2 The Structure of Fuzzy Physics-Informed Neural Network Hybrid Model

We focus on Single Input Single Output (SISO) processes with the control signal (the manipulated variable) u , the output signal (the controlled variable) y and n_x state variables $x = [x_1 \dots x_{n_x}]^T$. The block diagram of the proposed PIHNN model structure is depicted in Fig. 1. It can be divided into three parts. The first is a series of n data-driven GRU neural network sub-models. Each of them is trained independently using one data set from a narrow neighbourhood of one of the operating points of the process; each set can have a different number of data samples. It means that we have n disjointed data sets. Input vectors of the GRU sub-models are represented by the vectors $\mathbf{X}_{\text{GRU}}^1(k), \dots, \mathbf{X}_{\text{GRU}}^n(k)$ (they may be equal), defined by $\mathbf{X}_{\text{GRU}}^i(k) = [u(k-1) \dots u(k-n_B)]^T$, where $i = 1, \dots, n$, while their outputs by the scalars $y_{\text{GRU}}^1(k), \dots, y_{\text{GRU}}^n(k)$; the symbol k stands for the current discrete sampling instant. Next, we have an FP model based on ordinary differential equations that describe the phenomena taking place in the process. The vector $\mathbf{X}_{\text{FP}}(k) = [x^T(k-1) u(k-1)]^T$ defines the input of the FP sub-model, and its output is given by the scalar $y_{\text{FP}}(k)$. Finally, the outputs of GRU neural network sub-models and the output of the FP model are fed as inputs of the Fuzzy Data Fusion Block (DF). The DF block, based on the current operating point of the process, defined by the vector $\mathbf{X}_{\text{DF}}(k) = u(k-1)$ or $\mathbf{X}_{\text{DF}}(k) = y(k)$, combines output signals of GRU and FP sub-models to minimise the model error of the overall PIHNN structure. The output of the whole hybrid model is denoted by $y_{\text{PIHNN}}(k)$.

Different structures of the DF block may be used. Nevertheless, in this work, we consider a fuzzy structure. Such a choice is justified since we assume that

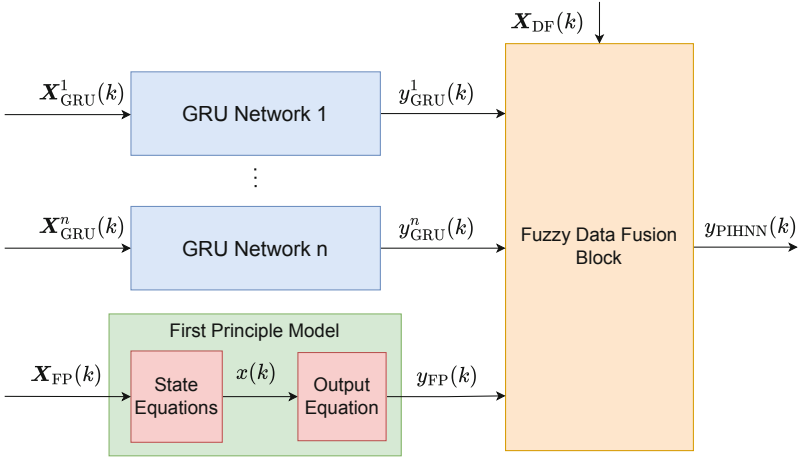


Fig. 1. Block diagram of the model structure

we have n independent GRU sub-models, each of which is trained for a specific operating point range and inactive outside these ranges. Similarly, we assume that the FP model is reliable around some chosen operating point. In such a case, the fuzzy DF block is a straightforward choice as it easily activates or deactivates the consecutive sub-models depending on the current operating conditions.

3 Simulation Results

To test the proposed PIHNN structure, we consider a neutralisation reactor benchmark [6]. The reactor is a SISO process with one input (the stream flow rate of NaOH, denoted as q_1) and one output (the pH value of the product). The sampling time is 10 sec. The process is highly nonlinear and is often used as a benchmark to test various control algorithms, e.g. [10]. During the experiments, we assume that data from the process are available in the form of two disjointed data sets, with measurements collected for $2.5 \leq \text{pH} \leq 4$ and $10 \leq \text{pH} \leq 15$. These data sets are used to train two independent GRU networks. We also have the FP model of the process in the form of differential equations. However, as in practice, the FP models can be inaccurate because some variables may not be measurable or we do not know model parameters precisely. In this work, we simulate such inaccuracies by changing the gain of the FP model to 90% of the original process's gain. All data sets have 5000 samples.

First, let us discuss the properties of the individual sub-models used. Two neural sub-models, GRU_1 and GRU_2 (each with $n_{\text{N}} = 7$ neurons and $n_{\text{B}}=2$ order of dynamics) have been trained using two distinct training data sets for different operating conditions. The first set contains measurements when the process output is $2.5 \leq \text{pH} \leq 4$, while the second set comprises measurements for $10 \leq \text{pH} \leq 15$. Both sub-models have the same parameters, i.e. seven neurons

and the second order of dynamics, which provide good modelling quality for the studied process, as described in [16]. The FP model is comprised of differential equations [6] with incorrect steady-state gain (90% of the original gain). To check the validity of all models, a validation data set covers the whole range of the process operating point, i.e., $2 \leq \text{pH} \leq 11$. Figure 2 shows 1000 samples of the validation data set vs outputs of two local GRU sub-models and the FP model with an incorrect gain. As expected, the output of the FP model is always smaller than the actual value of the output. GRU₁ and GRU₂ neural sub-models provide good modelling quality only for two distinct data ranges, i.e., the first for low pH values and the second one for high values of pH, respectively. The neural sub-models give huge errors for the data range for which they have not been trained.

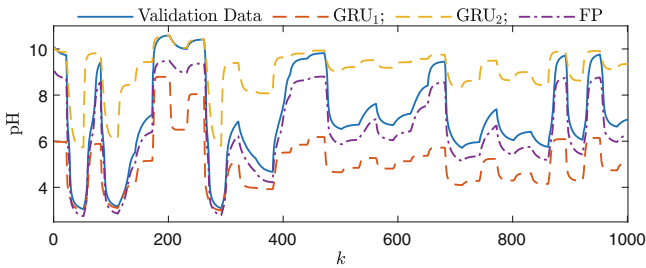


Fig. 2. 1000 samples of the validation data set vs outputs of two local GRU sub-models and the FP model with an incorrect gain.

Next, it is worth moving on to the conceptually most straightforward way to combine the outputs of all three sub-models. Figure 3 prefaces how the arithmetic average of the outputs of all sub-models looks against the validation data. Such an approach, despite its simplicity, has an obvious drawback since, for the areas of process operation where individual neural sub-models perform correctly, the averaged model has significant errors. Mediocre accuracy is obtained only for $4 \leq \text{pH} \leq 10$, but it is far below expectations.

Next, we consider several fuzzy PIHNN models with different Membership Functions (MFs). We considered trapezoidal

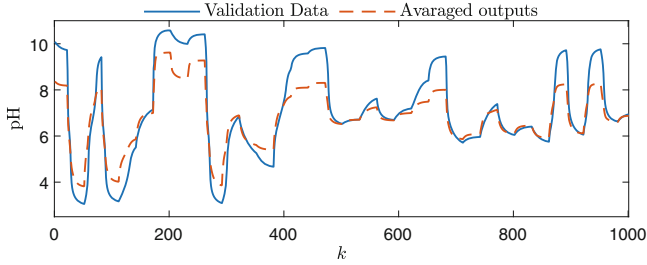
$$\mu(\text{pH}, \alpha, \beta, \gamma, \delta) = \max \left(\min \left(\frac{\text{pH} - \alpha}{\beta - \alpha}, \frac{\delta - \text{pH}}{\delta - \gamma}, 1 \right), 0 \right) \quad (1)$$

sigmoid

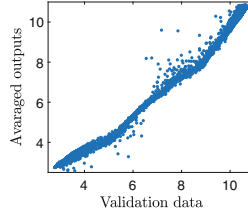
$$\mu(\text{pH}, \epsilon_1, \zeta_1, \epsilon_2, \zeta_2) = \frac{1}{1 + e^{-\epsilon_1(\text{pH} - \zeta_1)}} \frac{1}{1 + e^{-\epsilon_2(\text{pH} - \zeta_2)}} \quad (2)$$

and Gauss MFs

$$\mu(\text{pH}, \sigma, \eta) = e^{-\frac{(\text{pH} - \eta)^2}{2\sigma^2}} \quad (3)$$



(a) 1000 samples of the validation data set vs the averaged model output.



(b) The relation between the validation data set vs the averaged model output.

Fig. 3. Modelling using the averaged structure.

where α , β , γ , δ , ϵ_1 , ζ_1 , ϵ_2 , ζ_2 , σ and η are parameters of the considered MFs. Initially, these parameters were chosen manually and later fine-tuned using an optimisation procedure to minimise the PIHNN model error. As a result, we have obtained six fuzzy PIHNN models:

1. PIHNN ver. 1 with trapezoidal MFs and parameters chosen manually;
2. PIHNN ver. 2 with trapezoidal MFs and optimised parameters;
3. PIHNN ver. 3 with sigmoid MFs and parameters chosen manually;
4. PIHNN ver. 4 with sigmoid MFs and optimised parameters;
5. PIHNN ver. 5 with Gauss MFs and parameters chosen manually;
6. PIHNN ver. 6 with Gauss MFs and optimised parameters.

Figure 4 presents manually chosen and optimised membership functions. Regarding initial MFs, one can see that the shape of trapezoidal and sigmoid functions has been chosen so that the GRU models are used when pH values are within the training data set of the neural models. When pH values are far from the data used in training, the FP model is used. In other cases, a combination of GRU and FP model outputs' is calculated. The shape of Gaussian functions is more troublesome to select, as their tails are wide. As a result, the output of the PIHNNs model ver. 5 is always a combination of outputs of each sub-model. As far as optimised MFs are considered, we can see that optimisation results differ in each case. For trapezoidal MFs, we observe that the left fuzzy set becomes narrower, and both other sets expand. On the other hand, the left set becomes narrower for sigmoid functions, the right set expands, and the middle one changes its slopes

only slightly. In contrast, the middle set expands significantly for Gaussian functions, while the others moderately change their shape.

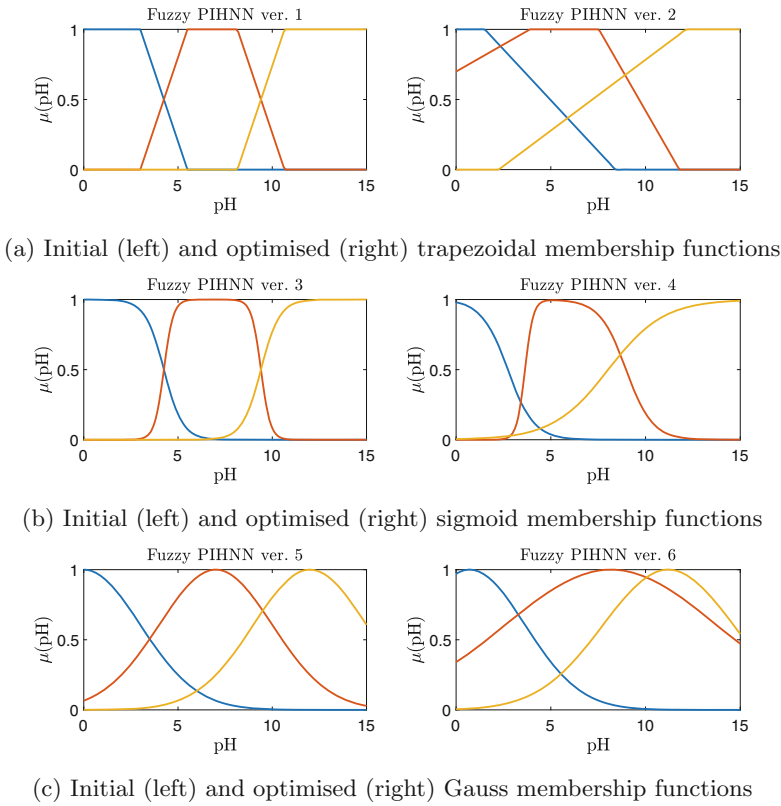
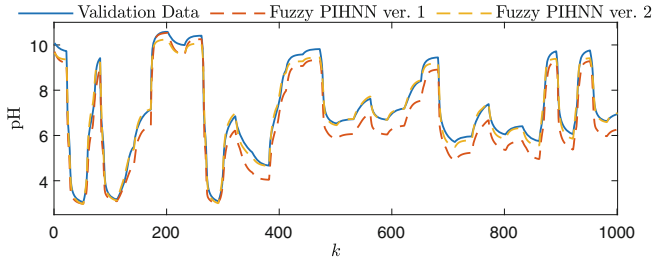


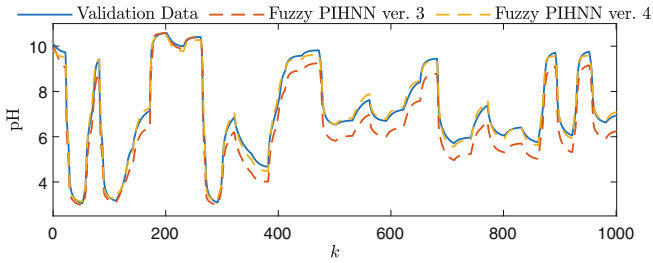
Fig. 4. Membership functions for considered fuzzy PIHNN models.

Figure 5 shows 1000 samples of the validation data set vs outputs of considered fuzzy PIHNN models. We can observe that PIHNN models ver. 1 and 3 (Figs. 5a and b, respectively) provide excellent modelling quality when pH values are at the extremes of the process operation area. However, the error increases when values of PH are in the region $4 \leq \text{pH} \leq 10$. On the other hand, the PIHNN model ver. 2 (Fig. 5a) is characterised by slightly worse modelling quality for large pH values. In contrast, the errors for $4 \leq \text{pH} \leq 10$ are considerably smaller in this case. The PIHNN model ver. 4 (Fig. 5b) provides excellent modelling quality for the whole data range. Finally, PIHNN models ver. 5 and 6, due to the wide shape of the MFs, have quite significant modelling errors when pH values are large.

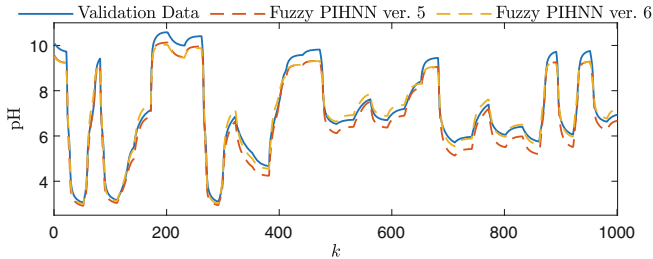
Figure 6 presents the relation between the validation data set and the outputs of all considered PIHNN models for the whole validation set, i.e., 5000 samples.



(a) 1000 samples of the validation data set vs the output of initial and optimised fuzzy PIHNN structures with trapezoidal MFs (PIHNN models ver. 1 and ver. 2).



(b) 1000 samples of the validation data set vs the output of initial and optimised fuzzy PIHNN structures with sigmoid MFs (PIHNN models ver. 3 and ver. 4).

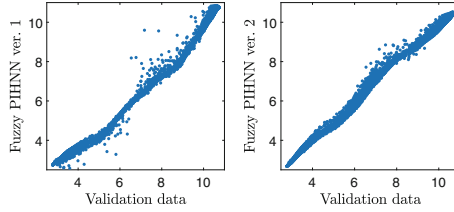


(c) 1000 samples of the validation data set vs the output of initial and optimised fuzzy PIHNN structures with Gauss MFs (PIHNN models ver. 5 and ver. 6).

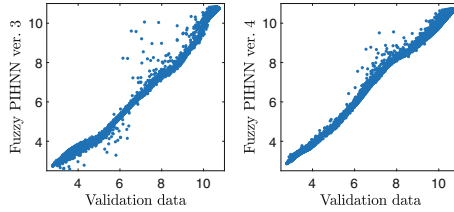
Fig. 5. 1000 samples of the validation data set vs the outputs of considered fuzzy PIHNN models.

One can observe (Figs. 6a and b) that PIHNN models ver. 1 and 3 frequently generate the output signal that differs significantly from the actual process value. For PIHNN models ver. 2 and 4, significant errors are less prevalent. In this context, PIHNN models ver. 5 and 6 work best (Fig. 6c); the output of these models is closest to the real process output for all data samples.

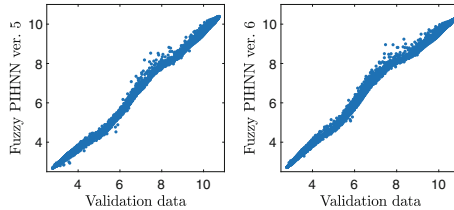
The Mean Squared Errors (MSE) for all PIHNN models and simple averaging of sub-model are shown in Table 1. The averaging approach has the largest error. The discussed PIHNN structure with simple trapezoidal or sigmoid MFs chosen manually reduced the model error almost three times. In the case of Gaussian



(a) The relation between the validation data set vs outputs of initial and optimised fuzzy PIHNN structures with trapezoidal MFs (PIHNN models ver. 1 and ver. 2).



(b) The relation between the validation data set vs outputs of initial and optimised fuzzy PIHNN structures with sigmoid MFs (PIHNN models ver. 3 and ver. 4).



(c) The relation between the validation data set vs outputs of initial and optimised fuzzy PIHNN structures with Gauss MFs (PIHNN models ver. 5 and ver. 6).

Fig. 6. The relation between the validation data set vs outputs of considered fuzzy PIHNN models.

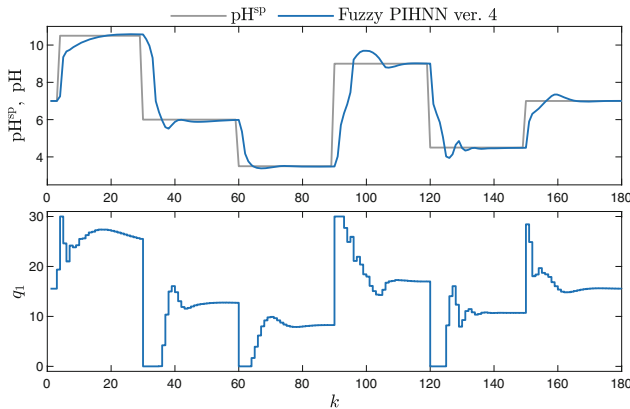
MFs, the error is even more minor. For all MFs studied, optimisation of the parameters of MFs results in significantly lower errors compared with errors of the corresponding initial model structures. Optimisation gives the best results for the sigmoid-shaped MFs. Other MFs, however, have only slightly larger errors.

Since the model PIHNN ver. 4 has the smallest error, it is used in the MPC algorithm. In each MPC sampling step, the prediction is determined from the model. The optimal sequence of the manipulated variable is computed using the Sequential Quadratic Programming (SQP) optimisation algorithm. Results of experiments are presented in Fig. 7. The MPC algorithm performs excellently when $\text{pH}^{\text{SP}} = 11$ and $\text{pH}^{\text{SP}} = 3.5$. There is no overshoot, no oscillations of the process output, and the settling time is short. This is because only GRU sub-models are active near those operating points; therefore, an accurate prediction is computed by the PINN model. For different values of pH^{SP} , the control quality is slightly worse as there is a small overshoot and oscillations of the output signal

Table 1. MSE errors for averaged and fuzzy PIHNN models with different MFs.

Model type	Avaraged	PIHNN		
		Trapezoidal MFs	Sigmoid MFs	Gauss MFs
Initial	164.33	57.49	61.91	29.32
Optimised	–	10.25	7.01	19.55

are observed for $\text{pH}^{\text{SP}} = 4.5$. In those areas, the combination of GRU and FP sub-models is used as prediction, computed by the fuzzy data fusion block. However, the control quality is still good, and the set-point is achieved fast.

**Fig. 7.** MPC simulation results for chosen fuzzy PIHNN best performing model.

4 Conclusions

The PIHNN model presented in this work relies on a series of black-box sub-models and a fundamental one. Each sub-models is trained independently for a different data range, while the fundamental model may have an error. All sub-models are combined using the fuzzy approach. The discussed model structure makes it possible to obtain low values of model errors for the whole data range. Additionally, the PIHNN model performs very well in MPC.

Acknowledgement. This research was financed by Warsaw University of Technology in the framework of the research project for the scientific discipline automatic control, electronics and electrical engineering.

References

1. Alexis, K., Nikolakopoulos, G., Tzes, A.: Switching model predictive attitude control for a quadrotor helicopter subject to atmospheric disturbances. *ISA Trans.* **19**, 1195–1207 (2011)
2. Alhajeri, M.S., Luo, J., Wu, Z., Albalawi, F., Christofides, P.D.: Process structure-based recurrent neural network modeling for predictive control: a comparative study. *Chem. Eng. Res. Des.* **179**, 77–89 (2022)
3. Assandri, A.D., de Prada, C., Rueda, A., Martínez, J.S.: Nonlinear parametric predictive temperature control of a distillation column. *Control. Eng. Pract.* **21**, 1795–1806 (2013)
4. Balla, K.M., Nørgaard, J.T., Bendtsen, J.D., Kallesøe, C.S.: Model predictive control using linearized radial basis function neural models for water distribution networks. In: 2019 IEEE Conference on Control Technology and Applications (CCTA), Hong Kong, pp. 368–373 (2019)
5. Camacho, E.F., Bordons, C.: *Model Predictive Control*. Springer, London (1999). <https://doi.org/10.1007/978-0-85729-398-5>
6. Gómez, J.C., Jutan, A., Baeyens, E.: Wiener model identification and predictive control of a pH neutralisation process. *Proceed. IEE Part D Control Theory Appl.* **151**, 329–338 (2004)
7. Hosen, M.A., Hussain, M.A., Mjalli, F.S.: Control of polystyrene batch reactors using neural network based model predictive control (NNMPC): an experimental investigation. *Control. Eng. Pract.* **19**, 454–467 (2011)
8. Ławryńczuk, Maciej: Computationally Efficient Model Predictive Control Algorithms. *SSDC*, vol. 3. Springer, Cham (2014). <https://doi.org/10.1007/978-3-319-04229-9>
9. Ławryńczuk, M.: Modelling and predictive control of a neutralisation reactor using sparse support vector machine wiener models. *Neurocomputing* **205**, 311–328 (2016)
10. Ławryńczuk, Maciej: *Nonlinear Predictive Control Using Wiener Models*. *SSDC*, vol. 389. Springer, Cham (2022). <https://doi.org/10.1007/978-3-030-83815-7>
11. Lima, P.F., Pereira, G.C., Mårtensson, J., Wahlberg, B.: Experimental validation of model predictive control stability for autonomous driving. *Control. Eng. Pract.* **81**, 244–255 (2018)
12. Roehrl, M.A., Runkler, T.A., Brandtstetter, V., Tokic, M., Obermayer, S.: Modeling system dynamics with physics-informed neural networks based on Lagrangian mechanics. *IFAC-PapersOnLine* **53**(2), 9195–9200 (2020). 21st IFAC World Congress
13. Schwedersky, B.B., Flesch, R.C.C., Dangui, H.A.S.: Practical nonlinear model predictive control algorithm for long short-term memory networks. *IFAC-PapersOnLine* **52**, 468–473 (2019)
14. Tatjewski, P.: *Advanced control of industrial processes, structures and algorithms*. Springer, London (2007). <https://doi.org/10.1007/978-1-84628-635-3>
15. Yang, L., Meng, X., Karniadakis, G.E.: B-PINNs: Bayesian physics-informed neural networks for forward and inverse PDE problems with noisy data. *J. Comput. Phys.* **425**, 109913 (2021)
16. Zarzycki, K., Ławryńczuk, M.: LSTM and GRU neural networks as models of dynamical processes used in predictive control: a comparison for two chemical reactors. *Sensors* **21**, 5625 (2021)



Evaluating Hydrodynamic Indices of the Underground Gas Storage Operation Based upon a Two-Phase Filtration Model

Ivan Sadovenko , Olexander Inkin , and Nataliia Dereviahina  

Dnipro University of Technology, 19 Dmytra Yavornytskoho Ave, Dnipro 49005, Ukraine
natali.derev@gmail.com

Abstract. The research objective is to develop and test mathematical model of gas storage in the layered aquifer with poorly permeable interlayer if plane-parallel and axisymmetric filtration takes place. The paper evaluates the gas-hydrodynamic operational indices of the underground gas storages within aquifers in the South East Ukraine. Comprehensive approach has been applied involving collection, systematization, and analysis of actual data on filtration and physicochemical properties of enclosing rocks impacting formation of natural and technogenic deposits as well as analytical and numerical methods to solve equations of the gas-water contact shift under different conditions. A gas-hydrodynamic model of underground gas storage within the nonuniform aquifer has been substantiated to calculate its cyclic operation in the three-layered seam taking into consideration crossflows through a poorly permeable stopping. The calculation results show significant impact of characteristics of the layered porous environment on the gas water contact transfer through certain seams. The derived new technique linearizing a system of differential equations to identify pressure within a reservoir is generalization of the earlier applied procedures with introduction of ‘boundary schemes’. The calculation results demonstrate significant impact of the layered porous environment on the gas water contact transfer through certain seams. The findings may be applied while making evaluations at the stage of gas storage design within aquifers.

Keywords: Aquifer · Gas Storage · Filtration · Gas Water Contact · Nonuniformity

1 Introduction

Along with the necessity to develop alternative energy sources, stable operation of fuel and energy complex of Ukraine depends heavily upon the reliable functioning of the unified gas supply system involving production facilities to mine, transport, store, and distribute gaseous hydrocarbons. A significant feature of the system is complete interconnection of its components expressed by changes in operation conditions of the system if working conditions of its certain object varies. In such a way, nonuniform gas consumption may result in its interrupted recovery while demanding the development of

underground gas storages (UGSs) within the deposits of aqueous rock as well as mathematical models able to calculate hydrodynamic indices of their operations under different geological and technological conditions [1–5].

The earlier considered [5, 6] hydrodynamic models of UGSs were obtained while assuming piston nature of water displacement with gas. Such a schematization of the process is popular and completely justified in many cases [7, 8]. Nevertheless, optimum ratio between buffer gas volume and active one, and determination of coefficients of gas saturation and average weighed pressure for different aquifer zones cannot be identified in terms of a piston problem formulation.

The problem has been solved partially under Buckley-Leverett theory; numerous papers concern it (for example, [9, 10]). The essential point is as follows: in this context, gas saturation distribution has been determined under constant initial conditions irrespective of solution for gas which makes it possible to simplify drastically the calculation procedure. However, the abovementioned complicates interpretation of the calculation results since the undefined pressure prevents from recalculation of gas amount within the seam to the normal conditions. In this connection, the paper objective is to substantiate the methods determining the basic hydrodynamic UGS indices in terms of its cyclic operation based upon a two-phase filtration model, and determination of the average weighed pressure and gas saturation within different storage zones. The peculiarity of the proposed gas-hydrodynamic model of an underground gas storage created in a heterogeneous aquifer is that it allows for the calculation of its cyclic operation in a three-layer formation, taking into account the interlayer flows through a low-permeability barrier.

2 Research Material and Methods

Dynamics of cyclic water displacement with gas is analyzed within a uniform infinite reservoir. A radial displacement case is considered. It is anticipated that in terms of degassing, mass discharge of gas $\rho a t G(t)$ is known and gas saturation at the well is constant. Gas is extracted until $R(t)$ front, having a high water saturation value, approaches the well.

It is also anticipated that the displacement process forms three typical zones (Fig. 1): 1st being a zone with high average gas saturation $\bar{\sigma}_1$ limited by a circle with $R(t)$ radius; 2nd being a zone with low average gas saturation $\bar{\sigma}_2 < \bar{\sigma}_1$; and 3rd being a zone inflated with pure water $\bar{\sigma}(x, y) = 0$. Since pressure within a high average gas saturation zone is almost equal to pressure within $R(t)$ front, we assume that pressure in the 1st zone depends only upon time. We also assume that within the 2nd and 3rd zones, pressure follows the equation of elastic liquid filtration.

Pressure distribution within the 2nd and 3rd zones is identified using a method of ‘fictive’ sources and drainages [11]. Thus, we have

$$P(r, t) = P_k + \frac{\mu_v}{4\pi kh} \int_0^t \frac{Q_0(\tau)}{(t - \tau)} \exp\left(-r^2/(4a(t - \tau))\right) d\tau, \quad (1)$$

where P_k is pressure at infinity; μ_v is water viscosity; k , h , and a are permeability, thickness, and piezoconductivity of a seam; and $Q_0(\tau)$ is specific rate of a ‘fictive’ source located in the central share of a seam.

Within the 1st boundary, $P_f(R(t), t)$ pressure and hence average weighed $\bar{P}(t)$ pressure is determined relying upon Eq. (1)

$$\bar{P}(r, t) = P_k + \frac{\mu_v}{4\pi kh} \int_0^t \frac{Q_0(\tau)}{(t-\tau)} \exp\left(-R^2(t)/(4a(t-\tau))\right) d\tau, \quad (2)$$

The specific rate of a ‘fictive’ source $Q_0(\tau)$ is selected in such a way to equalize at the $r = R(t)$ front consumption on the left and right of the boundary. Within the 1st zone, the total $Q(t)$ consumption is constant along the whole area (under the assumption on incompressibility of phases being filtered) inclusive of $R(t)$ boundary. From the 2nd zone, $Q(t)$ consumption may be derived using Eq. (1). Hence, to identify specific rate of the ‘fictive’ source, we have following integral equation

$$Q(t) = 2\pi R(t)h \left(-\frac{k}{\mu_v} \frac{\partial P}{\partial r} \right)_{r=R(t)} = \frac{R^2(t)}{4a} \int_0^t \frac{Q_0(\tau) e^{-\frac{R^2(t)}{4a(t-\tau)}}}{(t-\tau)^2} d\tau. \quad (3)$$

A law of $R(t)$ front advance is known from saturation solution [11]

$$R^2(t) = \frac{f'(\sigma^+)}{\pi mh} \int_0^t Q(t) dt, \quad (4)$$

where $f'(\sigma^+)$ is the derived Backley-Leverett function applied for gas saturation within the front; and m is a seam porosity.

Equation of gas balance is used to close (1)–(4) system.

$$P_t \int_0^t G(t) dt = \bar{P}(t) \pi R^2(t) mh \bar{\sigma}_1 + \int_{R(t)}^{R^*} P(r, t) 2\pi h r m \bar{\sigma}_2 dr \quad (5)$$

$P(r, t)$ function, located under integral in expression (5), is determined from ratio (1). However, in view of the 2nd limitedness, it is possible to apply simpler logarithmic pressure distribution corresponding to incompressible liquid filtration

$$P(r, t) = \tilde{P}(t) - \frac{Q(t) \mu_v}{2\pi kh} \ln \frac{r}{R(t)}, \quad (r > R(t)). \quad (6)$$

Having inserted ratio (6) into Eq. (5) and calculated the integral in the right side, we will obtain

$$P_t \int_0^t G(t) dt = \tilde{P}(t) \int_0^t Q(t) dt - \frac{Q(t) \bar{\sigma}_2 \mu_v}{K} \left(\frac{R^2}{2} \ln \left(\frac{R^{*2}}{R(t)^2} - \frac{1}{4} (R^{*2} - R^2(t)) \right) \right), \quad (7)$$

where $G(t)$ is consumption of gas injected into the seam under the standard conditions.

It should be mentioned that the simplification, connected with substitution of expression for $P(r, t)$, is not of principal nature; thus, ratio (1) can be used for Eq. (5) during numerical implementation.

Introduce dimensionless variables

$$x = \frac{t}{T}; \tilde{P} = \frac{\tilde{P}}{P_k}; \zeta = \frac{\tau}{T}; q = \frac{Q \mu_v}{P_k h k}; q_0 = \frac{Q_0 \mu_v}{P_k h k}; \alpha = \frac{R^2}{4aT}, \quad (8)$$

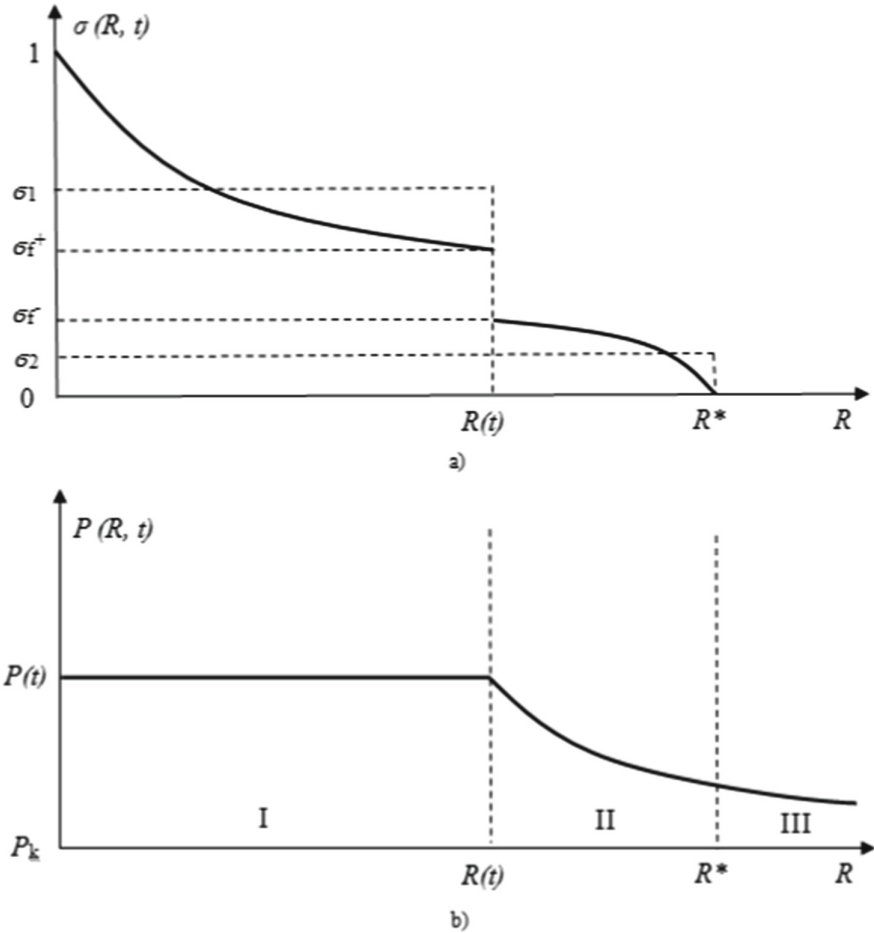


Fig. 1. Diagrams to calculate gas saturation (a) and pressure (b) within a horizontal aquifer

where T is typical process life; as a rule, it is equal to a year.

Owing to the use of the introduced variables, (2)–(4), and (7) expression will look like

$$\tilde{p}(x) = 1 + \frac{1}{4\pi} \int_0^x \frac{q_0(\zeta) e^{-\frac{\alpha(x)}{x-\zeta}}}{x-\zeta} d\zeta, \tag{9}$$

$$q(x) = \alpha(x) \int_0^x \frac{q_0(\zeta) e^{-\frac{\alpha(x)}{x-\zeta}}}{(x-\zeta)^2} d\zeta, \tag{10}$$

$$\alpha(x) = \beta f'(\sigma^+) \int_0^x q(x) dx, \tag{11}$$

$$\gamma W = \tilde{p}(x) \int_0^x q(x) dx - \frac{q(x)}{4\pi\beta} \left(\alpha^* \ln \frac{\alpha^*}{\alpha} - (\alpha^* - \alpha) \right), \tag{12}$$

Assume that injection and withdrawal follow a harmonic law.

(9)–(13) ratios are the closed system of integral equations to identify $\tilde{p}(x)$, $q(x)$, $q_0(x)$, and $\alpha(x)$. In this context, $\bar{\sigma}_1$, $\bar{\sigma}_2$, and σ^+ values are determined along with solving a problem for saturation [11].

Then single out and consider separately three specific stages of UGS operation: initial gas injection into the undisturbed aquifer; extraction; and gas injection into the seam during a random cycle of UGS operation.

Initial gas injection into the undisturbed aquifer. If gas is injected into the undisturbed aquifer then gas saturation distribution is represented by means of the known Backley-Leverett solution [12]; zone 2 (Fig. 1) is not available, $R^* = R(t)$. In this regard, the saturation jump $R(t)$ is defined using the ratio

$$R^2(t) = \frac{f'(\sigma^+)}{2\pi kh} \int_0^t Q(\tau) d\tau. \tag{13}$$

Gas saturation σ^+ within $R(t)$ front is determined by means of the transcendental equation solution

$$\frac{f(\sigma^+)}{\sigma^+} = f'(\sigma^+) \tag{14}$$

Average gas saturation is identified from the ratio

$$\bar{\sigma}_1 = \frac{1}{f'(\sigma^+)} \tag{15}$$

Assume for the numerical (9)–(13) system that within $X_{j-1} \leq X \leq X_{j-1} + \Delta X_j$ moment, $q(x)$ and $q_0(x)$ functions are constant. Subsequently, (9)–(13) equations for the specified time interval may be represented as follows

$$q_j(X) = \frac{\gamma w(X) - \dot{p}_j(x) \sum_{i=1}^{j-1} q_i \Delta x_i}{\dot{p}_j(x) \Delta X_j}; \tag{16}$$

$$\alpha_j(x) = \beta f'(\sigma^+) \sum_{i=1}^j q_i \Delta x_i; \tag{17}$$

$$q_0(x) = (q_j - \sum_{i=1}^{j-1} q_{0i} \left(e^{-\frac{\alpha_j}{x-x_{i-1}}} - e^{-\frac{\alpha_j}{x-x_i}} \right) e^{\frac{\alpha_j}{\Delta x_j}} \tag{18}$$

$$\dot{p}_j(x) = 1 + \frac{1}{4\pi} \sum_{i=1}^{j-1} q_{0i} \left(-E_i \left(-\frac{\alpha_j}{x-x_{i-1}} \right) + E_i \left(-\frac{\alpha_j}{x-x_i} \right) \right). \tag{19}$$

$W(x)$ function is given by the expression

$$W(x) = 0.5(1 + \cos(2\pi x)). \tag{20}$$

Analytical model for the case looks like: $W(x)$ is the initial value to solve (16)–(19) system. The value is determined through ratio (20). While inserting $W(x)$ into (16) and setting $\hat{p}_j^*(x)$ value, to a first approximation, it may be specified as that one being equal to \hat{p}_j-1 . Identify $q_j(x)$ value. (17)–(19) ratios help define successively α_j , q_{oj} , and \hat{p}_j values. In general, the latter does not coincide with \hat{p}_j^* . New approximation of \hat{p}_j^* is selected; the iteration process continues until \hat{p}_j^* matches \hat{p}_j with the specified E accuracy. The final $\hat{p}_j(x)$ value, calculated for the time interval, also defines $q_j(x)$ and $\alpha_j(x)$ values corresponding to it.

Gas extraction during a random cycle. As it has been mentioned above, according to the accepted planning, gas is extracted until saturation jump with a coordinate nears a well placed in the central share of the seam. In this context, volume of the gas, extracted from Q_k^* seam, may be defined as well as normalized to the reservoir conditions. Then, relying upon the specified harmonic selection law, it becomes possible to identify gas consumption normalized to the reservoir conditions

$$q(x) = - \frac{\pi Q_H^*}{T} \sin (2\pi (x - xN)) \tag{21}$$

In this case, the equation takes the form

$$\alpha_j(x) = \alpha^* - \beta f'(1 - 6f) \sum_{i=N}^j q_i \Delta x_i \tag{22}$$

where α^* is the maximum $\alpha(x)$ value achieved during gas injection; N is number of xN time moment corresponding to the extraction start; and σf is front value of gas saturation in terms of k th extraction.

Since $q(x)$ value is entered by means of (21) then $\alpha_j(x)$, $q_{oj}(x)$, and $\hat{p}_j(x)$ values may also be identified through direct computation using formulas (17), (19), and (22); $W_j(x)$ value is defined using the formula

$$W_j(x) = \frac{1}{\gamma} (\hat{p}_j(x) \sum_{i=1}^j q_i \Delta x_i - \frac{q_i}{4\pi\beta} \left(\alpha^* \ln \frac{\alpha^*}{\alpha} - \alpha^* + \alpha \right)). \tag{23}$$

In such a way, while extracting, W_j is not the initial (specified) value. It is determined during the problem solving. The abovementioned depends upon the selected operational schedule of UGS. In the context of the schedule, gas extraction is maximum possible and volume of gas, remained in the seam, is minimal.

Under reservoir conditions Q^* , front saturation value σf as well as the extracted gas volume is determined through the solution analysis for saturation [13].

If the extraction takes place within a random k th cycle, then solution for saturation is divided into two cases:

1. In terms of $0 < \bar{\sigma}_2 < \sigma n$ (where σn is gas saturation corresponding to Backley-Leverett bending point function) case, volume of gas Q_k^* , extracted from a seam during k th cycle, is identified using the expression

$$Q_k^* = \beta \sum_{i=1}^N q_i \Delta x_i - \alpha^* \left(\bar{\sigma}'_2 - \frac{f(\bar{\sigma}'_2)}{f(\bar{\sigma}'_2)} \right) \tag{24}$$

2. If $\sigma_n < \bar{\sigma}_2 < 1$ then Q^* determination should involve gas saturation assessment within of front by means of the transcendental equation solving

$$\frac{f(\bar{\sigma}_2) - f(\sigma_f)}{\bar{\sigma}_2 - \sigma_f} = f'(\sigma_f) \quad (25)$$

While applying the determined σ_f value, the following is defined

$$Q_k^* = \beta \sum_{i=1}^N q_i \Delta x_i - \alpha^* (\sigma_f - \frac{f(\sigma_f)}{f'(\sigma_f)}) \quad (26)$$

At the end of the extraction, average gas saturation value $\bar{\sigma}_2$ is defined with the help of the formula

$$\bar{\sigma}_2 = \bar{\sigma} - \frac{Q^*}{\alpha^*} \quad (27)$$

Moreover, it is taken up as the initial distribution at the start of following injection (Fig. 2 c, d, and e).

Gas injection within a random $k + 1$ st cycle. Within a random cycle, gas injection into UGS differs from its initial injection in the fact that there is some gas saturation distribution in the seam; for the case, it is substituted for a constant $\bar{\sigma}_2$ value. During injection, the seam demonstrates two gas saturation jumps with $R(t)$ and R^* coordinates as well as corresponding dimensionless variables $\alpha(x)$ and α^* (Fig. 2, f).

In this case, expression (16), determining mass balance in the UGS, will take the form

$$q_j(x) = \frac{\gamma W_j(x) - \bar{p}_j(x) \sum_{i=1}^{j-1} q_i \Delta x_i}{\bar{p}_j \Delta x_j - \left(\frac{\bar{\sigma}_2}{4\pi\beta}\right) \left(\alpha^* \ln\left(\frac{\alpha^*}{\alpha_j}\right) - \alpha^* + \alpha_j\right)} \quad (28)$$

The unknowns q_0 , α_j , and \bar{p}_j are identified from (17)–(19) ratios. (1)–(19), and (28) system is solved similarly to the initial injection case. Law of $W_j(x)$ variation is taken up as follows

$$W_j(x) = W_0 + 0.5W_k + 1(1 - \cos(2\pi(-N))) \quad (29)$$

where W_0 is amount of gas in the seam before previous extraction is over; $W_k + 1$ is amount of gas injected in the seam during active injecting; and x_{N+1} is starting point of the active injective.

If gas is injected within a random cycle, the saturation solution depends upon the injected gas volume (under the reservoir conditions)

$$Q_3 = \int_{x_{N+1}}^{x_{N+2}} q(x) dx. \quad (30)$$

Having determined the value of front gas saturation of from the ratio

$$\frac{f(\sigma_f) - f(\bar{\sigma}_2)}{\sigma_f - \bar{\sigma}_1} = f'(\sigma_f), \quad (31)$$

and having identified ‘critical’ value of the injected gas volume

$$Q_{cr} = \frac{\alpha^* \bar{\sigma}_2}{\beta (\bar{\sigma}_2 f(\sigma_f) - f(\bar{\sigma}_2))}, \quad (32)$$

consider two probable solution alternatives:

1. Assume that Q_3 value is less than critical volume Q_{cr} ; then motion of two saturation jumps takes place within the seam. In this context, before injection is over, maximum zone the maximum zone with gas α_{N+2}^* , is (Fig. 2, f)

$$\alpha_{N+2}^* = \alpha_{N+1}^* + \beta Q_3 f'(\bar{\sigma}_2) / \bar{\sigma}_2. \quad (33)$$

Amount of gas within the seam as well as average gas saturation until injection is over will be defined as follows

$$Q_{N+2} = Q_{N+1} + Q_3, \quad (34)$$

$$\bar{\sigma}_1 = \beta Q_{N+2} / \alpha_{N+2}^*. \quad (35)$$

of value, being a part of Eq. (18), is defined from (31) ratio.

2. If $Q_3 > Q_{cr}$ then frontal gas saturation value is identified by solving the equation

$$\beta (\sigma_f f^l(\sigma_f) - f(\sigma_f) + 1) Q_3 - \bar{\sigma}_2 \alpha_{N+1}^* = \beta Q_3 \quad (36)$$

In addition, geometry of zone with gas before injection is over α_{N+2}^* is defined from the ratio

$$\alpha_{N+2}^* = \beta Q_3 f^l(\sigma_f) \quad (37)$$

Q_{N+2} and $\bar{\sigma}_1$ values are determined from (19) and (35) expressions.

Case two solving for injection arises if one saturation jump is in the seam (Fig. 2, g). It should be mentioned that the initial injecting stage will always involve case one; consequently, if a back edge $\alpha(x)$ nears and passes α^* ($Q_3 > Q_{cr}$) front case two may happen (Fig. 2, f, g).

The considered algorithm to solve the formulated problem has been represented based upon approximate solution for saturation [9]. It is possible to examine similar solution algorithm for the case when saturation is solved relying upon accurate statement [16]. Comparative analysis of two solutions, performed on the basis the processed calculation results, has shown their good agreement for the first six operational UGS cycles. Computation for more prolonged period, based upon a model with averaging, results in significant errors. However, simple implementation and short period, required

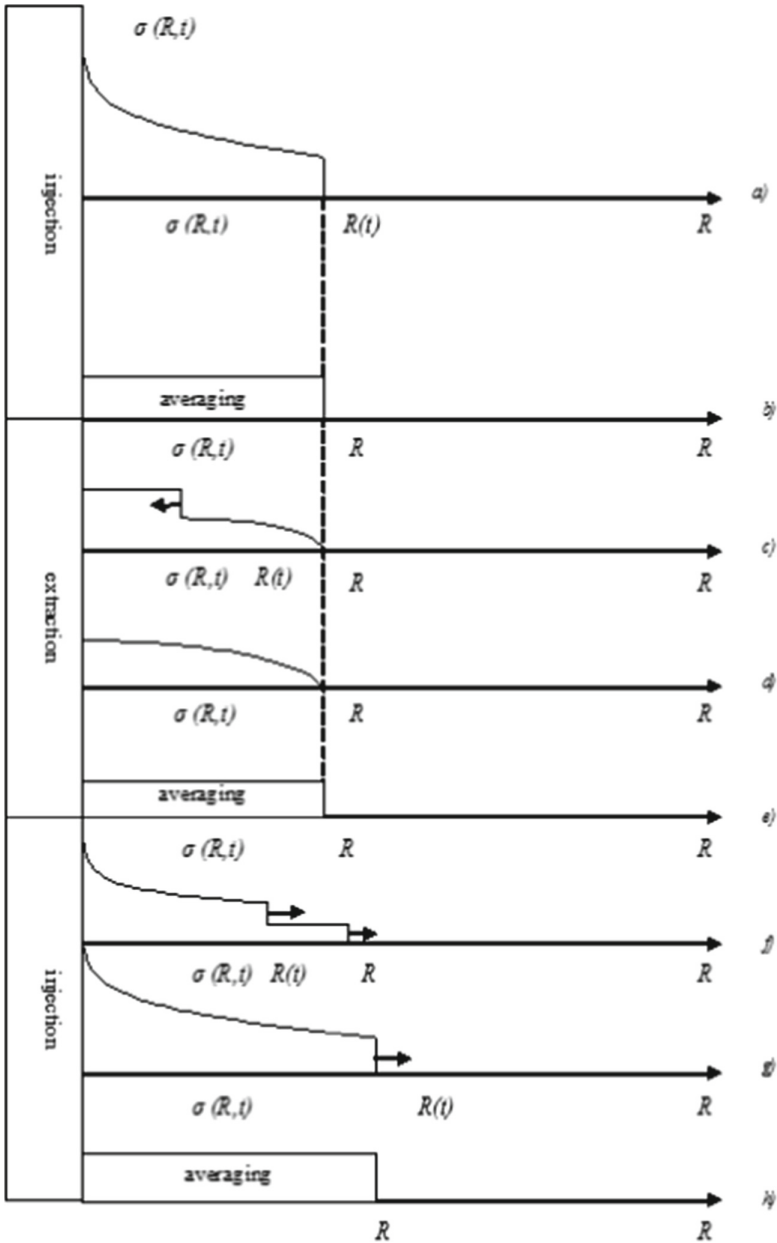


Fig. 2. On the calculation of gas saturation under cyclic UGS operation

to make the calculations, help recommend the methods while making multivariant computations of a gas storage transfer to cyclic operation. Essentially, this is a new method of linearizing a system of differential equations to determine pressures in a reservoir, which generalizes previously used methods. This method introduces “boundary schemes” as

mentioned above. The results of calculations based on the new methodology demonstrate a significant impact of the characteristics of the layered porous medium on the movement of the gas-water contact through individual layers.

3 Results and Their Analysis

The represented algorithm has been implemented in the software environment Mapple for a hypothetical case. The seam parameters were specified as follows: $\kappa = 10\text{--}12 \text{ m}^2$ being permeability; $m = 0.2$ being porosity; $a = 1 \text{ m}^2/\text{s}$ being piezoconductivity coefficient; $h = 10 \text{ m}$ being seam thickness; $\mu_v = 10^{-3} \text{ Pa}\cdot\text{s}$ being formation water viscosity; $P_\kappa = 9.8 \text{ MPa}$ being pressure within the undisturbed seam boundary; and period of the complete operational UGS cycle being $T = 1 \text{ year} = 3.15 \cdot 10^7 \text{ s}$ (four three month periods: injection – idle time – extraction – idle time). The calculations involved the idea that one and the same gas mass G_3 is injected into the seam during any period. Mass of the extracted Gext gas was defined during the solution.

Phase penetrations for gas $k_g(\sigma)$ and water $k_w(\sigma)$, involved by Backley-Leverett functions, i.e. $f(\sigma)$

$$f(\sigma) = \frac{k_g \mu_g}{k_g \mu_g + k_w \mu}$$

were assumed as follows according to paper [10]

$$\begin{cases} k_g(\sigma) = \left(\frac{\sigma-0.1}{0.9}\right)^{3.5} (4 - 3\sigma), & (0.1 \leq \sigma \leq 1) \\ k_w(\sigma) = \left(\frac{0.8-\sigma}{0.8}\right)^{3.5}, & (0 \leq \sigma \leq 0.8) \end{cases}$$

where σ is gas saturation.

Figure 3 shows the calculation results. Curve I is dimensionless gas consumption under the formation conditions $q(\tau)$; II is change in average weighed pressure in the seam (τ); III is change in space of pores with gas having high average gas saturation $\alpha(\tau)$; and IV is volume of gas in the seam normalized to the standard conditions $W(\tau)$. Analysis of the data explains that pressure in UGS during the operation transfers to a cyclic mode rather rapidly; moreover, amplitude changes in terms of pressure variations during different cycles are not higher than several percent. In this context, formation pressure excess over a reservoir boundary while gas injecting is 5% ($P = 10.3 \text{ MPa}$). Subsequently, when following idle period is over the pressure equalizes ($P = P_\kappa$); in the period of gas extraction, formation pressure is 3–5% less than its boundary pressure ($P = 9.5\text{--}9.3 \text{ MPa}$) depending upon the operational UGS cycle.

Dimensionless gas consumption under the formation conditions, being the ratio between product of its viscosity consumption and boundary pressure product per seam thickness and permeability, also increases up to 0.13 while injecting. It is almost matches $1100 \text{ m}^3/\text{day}$ gas consumption. During following idle time, gas consumption decreases vanishing to the period end. By the end of extraction stage, gas flow rate increases annually. It was 0.08 ($-678 \text{ m}^3/\text{day}$) in the first year; 0.11 ($-932 \text{ m}^3/\text{day}$) in the second

year; 0.13 (-1100 m³/day); and 0.14 (-1185 m³/day). It should be mentioned that after gas extraction and following idle time, its consumption was equal to zero again.

By the end of injection period, dimensionless space of pores with gas as well as during following idle time increased cycle by cycle. It was 2.5 in the first year; 4 in the second year; 4.8 in the third year; and 5.2 in the fourth year. The values correspond to a radius of a reservoir zone differing in high gas saturation, i.e. 17.7; 22.4; 24.5; and 25.6 km. In this regard, values of gas volume within the seam normalized to the standard conditions correspond to 1; 1.5; 1.8; and 2.1, respectively. It is also possible to mention a phase shift between $W(\tau)$ and $q(\tau)$ arising owing to the availability of elastic zone III (Fig. 1) with the formation liquid being contracted.

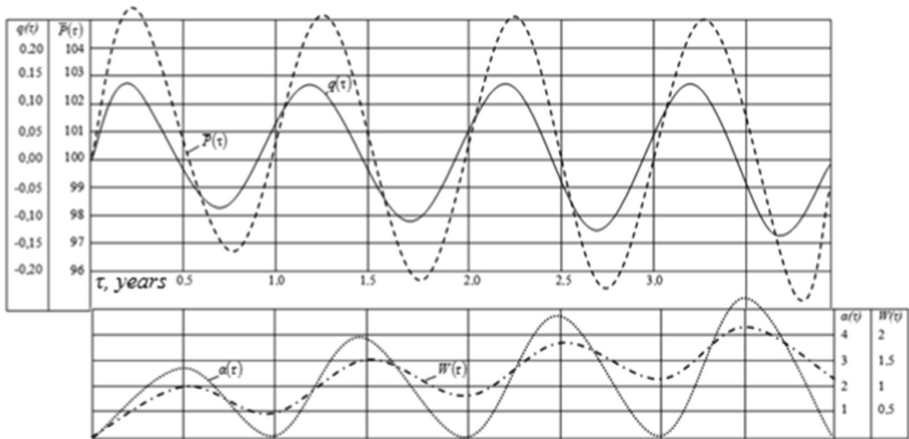


Fig. 3. Calculation example of hydrodynamic indices of UGS operation in terms of a cyclic mode

4 Conclusion

Numerical hydrodynamic model of underground gas storage within a horizontal aquifer has been developed; the model takes into consideration two-phase nature of liquid and gas filtration. Processing of the obtained results has made it possible to substantiate the approximate method calculating formation gas volume, consumption, and pressure at different stages of storage development as well as during its operational cycles. According to the calculation results in the software environment Maple for a hypothetic case, it has been identified that gas pressure in the storage transfers to a cyclic mode rather quickly under a minor (i.e. several percent) amplitude change during different cycles. In this context, formation pressure excess over a reservoir boundary while gas injecting is 5%; in the period of gas extraction, formation pressure is 3–5% less than its boundary pressure. In the injection period, gas consumption is almost constant; in turn, its flow rate during extraction increases year by year. In addition, space of pores with high average gas saturation also increases annually.

The proposed methods calculating the basic hydrodynamic UGS parameters in aquifers make it possible to identify the optimum ratio between buffer gas volume and

active one in the storage as well as determine the fundamental technical and economic indicators of its operation at the design stage. The abovementioned may be applied to make business plans and investment proposals concerning seasonal accumulation of gaseous hydrocarbons in the natural environment. Further research is expedient to test adequacy of the developed methods while comparing the obtained calculation results with actual operational data of the operating UGSs.

References

1. Sobczyk, W., Perny, K. C. I., Sobczyk, Eugeniusz, J.: Assessing the real risk of mining industry environmental impact. *J. Polish Mineral Eng. Soc.* **1**, 33–41 (2021). http://www.potopk.com.pl/Full_text/2021_v1_full/IM%201-2021-a5.pdf
2. Sobczyk, W., Sobczyk, Eugeniusz J. Varying the energy mix in the EU-28 and in Poland as a step towards sustainable development.. *Energies* **14** (5/1502), 1–19 (2021). <https://www.mdpi.com/1996-1073/14/5/1502>
3. Pysmennyi, S., Fedko, M., Chukharev, S., Kyelgyenbai, K., Anastasov, D.: Technology for mining of complex-structured bodies of stable and unstable ores. In: *IOP Conference Series: Earth and Environmental Science*, 970(1), p. 012040 (2022). <https://doi.org/10.1088/1755-1315/970/1/012040>
4. Pysmennyi, S., Peremetchyk, A., Chukharev, S., Anastasov, D., Tomiczek, K.: The mining and geometrical methodology for estimating of mineral deposits. *IOP Conference Series: Earth and Environmental Science*, 1049(1), 012029 (2022). <https://doi.org/10.1088/1755-1315/1049/1/012029>
5. Inkin, O., Derevyagina, N., Khrypilyvets, Y.: Modeling of gas storage performance in massive aquifers. *Physical and technical problems of mining. Institute of Physics of Mining Processes of the National Academy of Sciences of Ukraine*, 22, 31–45 (2020). (In Ukrainian)
6. Inkin, O., Tishkov, V., Dereviahina, N., Sotskov, V.: Integrated analysis of geofiltrational parameters in the context of underground coal gasification relying upon calculations and modelling. *Ukrainian School Mining Eng.* **60**, 1–9 (2018). <https://doi.org/10.1051/e3sconf/20186000035>
7. Voitenko, Yu., Vapnichna, V., Voitenko, O.: On the destruction and prefracturing of solid rocks under blasting in formation conditions. *Geoengineering* **7**, 7–16 (2022). <https://doi.org/10.20535/2707-2096.7.2022.267555> (In Ukrainian)
8. Zheltov, Yu. T.: *Mechanics of the oil and gas reservoir*. Nedra. (1975)
9. *Mining encyclopedia* / Edited by V.S. Biletskyi. Donetsk, Donbas (2004). (In Ukrainian)
10. Arkhipova, L.: The concept of ecological safety of basin systems of oil and gas production areas. *Ecological Secur. Balanced Resource Use* **2**(6), 67–71 (2012). (In Ukrainian)
11. Collins, R.E. T.: *Flow of fluids through porous materials*. United States (1976)
12. Bugay, Yu., Globa, V., Nagorny, V., & Vengertsev, Yu. T.: *Construction of oil depots and gas storage facilities*. Kyiv, VIPOL (2000). (In Ukrainian)
13. Sukhin, E. T.: *Elements of creation, formation and operation of underground gas storage facilities*. Kyiv, PNNV (2004). (In Ukrainian)
14. Boyko, V.S. T.: *Underground hydromechanics*. Kyiv, ISDO (1995). (In Ukrainian)



New Metrics of Fault Distinguishability

Jan Maciej Kościelny and Michał Bartys^(✉)

Institute of Automatic Control and Robotics, Warsaw University of Technology,
Boboli 8, 02-525 Warsaw, USA
{jan.koscielny,michal.bartys}@pw.edu.pl

Abstract. This paper addresses the problem of assessing fault distinguishability of faults. Fault distinguishability is understood as the ability of a diagnostic system to isolate faults. The objective of diagnosis is the early detection and accurate isolation of the faults that have occurred. The accuracy of diagnosis depends on the fault distinguishability achieved. The problem of assessing the fault distinguishability is therefore extremely important. This provides an opportunity to compare the effectiveness and assess the quality of different solutions of a diagnostic system. This paper proposes two new metrics for fault distinguishability assessment, providing the possibility to evaluate and compare the performance of fault isolation carried out based on both signatures or symptoms with both binary or trinary diagnostic signal values. The proposed metrics allow analysis of fault distinguishability for all diagnostic inference approaches based on Fault Signature Matrix as well on Fault Isolation System.

Keywords: fault distinguishability metrics · fault isolation · process diagnosis · diagnostic accuracy · diagnostic inference

1 Introduction

Fault distinguishability is understood as the ability of differentiating faults in result of inference based on outputs of diagnostic tests [13] and refers to the accuracy of diagnosis. The measure of diagnostic accuracy is the reciprocal of the number of potential faults indicated in the diagnosis [9]. Thus, the smaller the number of faults indicated in the diagnosis, the more accurate it is. Hence, increasing the distinguishability of faults consequently leads to an increase in the accuracy of diagnosis. A sufficiently accurate and reliable diagnosis is a prerequisite for taking effective protective actions. The distinguishability of faults therefore has a significant impact on process safety. Accurate isolation of faults that have occurred also makes it possible to immediately initiate repair or replacement actions of the defective process component(s). Clearly, this results in an increase in the functional safety and reliability indices of the entire system.

The distinguishability of faults depends on:

- the set of measurements used for diagnostic purposes;
- the set of diagnostic tests;
- the representation of the outputs of diagnostic tests (discrete or continuous);
- knowledge of the timed sequence in which the symptoms of faults occur.

Ensuring the required fault distinguishability is one of the most important issues by the design of a diagnostic system. In order to assess the achievable fault distinguishability for comparability analysis of different diagnostic approaches it is necessary to define measures of distinguishability.

Distinguishability of faults and its measures have often been defined in respect to the diagnostic method used, representation of diagnostic signals, and the form of the fault-diagnostic signal values relationship. However, it should be stressed that there is no so far any universal metric to assess and compare performance of any diagnostic system in terms of the achievable fault distinguishability.

This work defines two new metrics that would enable the assessment of fault distinguishability of diagnostic systems using different diagnosis approaches that are based on binary and multivalued diagnostic signals, and on the knowledge of the timed sequence of symptoms. However, it should be highlighted, that the proposed metrics do not apply for assessing fault classification approaches as well as fault isolation methods that take into account the impact of faults on the residuals.

In fact, the new metrics are applicable for the assessing of the majority of approaches used in the practice of diagnosing complex dynamic systems. These approaches are mostly based on expert knowledge of the fault-diagnostic signal values relationship. It is due to the impracticability of the approaches based of continuous diagnostic signals and fault classification approaches, because of their requirements regarding availability of the datasets representing all states with faults of diagnosed system.

2 Preliminaries

The automatic diagnosis is carried out based on the current values of the diagnostic signals generated by partial models of the diagnosed system and the knowledge of the fault-symptoms relationship formulated by experts in the design phase of the diagnostic system [11]. Optionally, the knowledge about the timed sequence of symptoms of given faults is also used. In order to make a diagnose, it is necessary to know the relationship between the set of faults F

$$F = \{f_k : k = 1, \dots, K\}, \quad (1)$$

and the set of values of diagnostic signals S

$$S = \{s_j : j = 1, \dots, J\}. \quad (2)$$

Two basic forms of notion for this relationship are used for approaches that use expert knowledge: Fault Signature Matrix (*FSM*) and Fault Isolation System (*FIS*). Both specify the possible values of diagnostic signals $s_j \in S$ for all faults $f_k \in F$.

The *FSM* as well as the *FIS* have the form of a table like structure in which the rows correspond to the diagnostic signals and the columns to the faults. The form of this table depends on the representation of the diagnostic signals:

- When binary evaluation of absolute values of residuals is used, the table takes the form of a *FSM* [3, 18, 24]. This form of the fault-symptom relationship, is also referred to as: structure of residual sets [4], boolean decision table [2], coding set [5, 17], effect of the faults on the residuals [1] or binary diagnostic matrix [7].
- When using multivalued residuals, the table takes the form of the *FIS* [7].

All other forms of this relationship, e.g. *if-then* rules, can be boiled down to *FSM*, or *FIS*, depending on the representation of diagnostic signals used [13]. The *FIS* has been defined in [9] as the following quadruplet:

$$FIS = \langle F, S, V_S, q \rangle, \quad (3)$$

where: F is the finite set of faults; S - the finite set of diagnostic signals; $V_S = \bigcup_{s_j \in S} V_j$ is the finite set of diagnostic signal values; V_j is the finite set of values of j^{th} diagnostic signal; q -mapping: $q : F \times S \rightarrow \phi(V_S)$, which associates each element of the Cartesian product $F \times S$ with a subset of the diagnostic signal values $q(f_k, s_j) \equiv V_{kj} \subset V_j$ which this signal can take, in the event of a fault f_k .

Let us further assume that the value of the diagnostic signal $s_j = 0$ corresponds to a fault free condition and that the non-zero values of s_j are symptoms of faults.

The fault signatures corresponding to the columns of *FSM* are referred to as simple signatures

$$V_{FSM}(f_k) = [v_{k1}, \dots, v_{kj}, \dots, v_{kJ}]^T. \quad (4)$$

In turn, the signatures of *FIS*, containing subsets of diagnostic signal values are referred to as complex signatures

$$V_{FIS}(f_k) = [V_{k1}, \dots, V_{kj}, \dots, V_{kJ}]^T. \quad (5)$$

In fact, the *FSM* is a special case of *FIS* if all values of all diagnostic signals are binary i.e. $V_S \in \{0, 1\}$.

The notion of relationship fault-diagnostic signals values in the form of *FSM* or *FIS* does not refer to the timed sequence of the symptoms of faults. However, the knowledge of this sequence can be used to distinguish between faults. In [13–15], are introduced the so-called elementary symptom sequences.

The elementary sequence $es_{j,p}(f_k) = \langle s_j, s_p \rangle$ defines the timed order of a pair of symptoms $\langle s_j, s_p \rangle$ of the same fault. This notation implies that after the occurrence of fault f_k , symptom s_j will occur prior to symptom s_p . Elementary

sequences allow conditionally distinguishing faults that are indistinguishable in *FIS* or *FSM*.

However, it should be noted that the knowledge of the timed symptom sequence is usually not complete. The elementary sequences usually can not be determined for all faults.

3 Definitions of Fault Distinguishability

Fault distinguishability is usually defined in the context of the diagnostic approach adopted and, in particular, in relation to the form of notion of the fault-diagnostic signals relationship. Definitions of indistinguishability of failure states and failure events (faults) were given by Kościelny in [8]. He defined *indistinguishable states (events) as being characterised by the identity of the results of the diagnostic tests*.

Gertler in [4] gave a definition of a weak distinguishability: *A structure is weakly isolating if all columns in the structure matrix are different and nonzero*. He defined also strong fault distinguishability (uni- and bidirectional). Strong distinguishability was intensively studied and analysed in [6].

Work [9] have shown the potential of increasing the distinguishability of faults by usage of multivalued diagnostic signals and *FIS*. In the case of multivalued diagnostic signals, the distinguishability cannot be determined unambiguously. It depends on the combination of diagnostic test results. Therefore, to tackle this problem, the concepts of unconditional and conditional fault distinguishability were proposed.

Definition 1. The faults $f_k, f_m \in F$ are indistinguishable (unconditionally indistinguishable) in *FIS* in respect to diagnostic signals $s_j \in S$ iff their signatures are identical.

$$f_k R_{UI} f_m \Leftrightarrow \forall_{s_j \in S} V_{jk} = V_{jm} \quad (6)$$

Definition 2. The faults $f_k, f_m \in F$ are conditionally distinguishable in *FIS* in respect to diagnostic signals $s_j \in S$ iff for each diagnostic signal, the intersection of sets of its values corresponding to faults f_k and f_m have a common elements and these faults are not unconditionally indistinguishable.

$$f_k R_{CI} f_m \Leftrightarrow \forall_{s_j \in S} V_{jk} \cap V_{jm} \neq \emptyset \wedge \exists_{s_j \in S} V_{jk} \neq V_{jm} \quad (7)$$

The conditional indistinguishability of faults means that there can present such values v_j of the diagnostic signals that satisfy the condition $\forall_{s_j \in S} v_j \in V_{jk} \cap V_{jm}$, at which the two given faults are indistinguishable. However, by other values of the diagnostic signals the same faults may be distinguishable. The following condition is then satisfied:

$$\exists_{s_j \in S} [v_j \in V_{jk} \wedge v_j \notin V_{jm}] \vee [v_j \notin V_{jk} \wedge v_j \in V_{jm}]. \quad (8)$$

Definition 3. The faults $f_k, f_m \in F$ are unconditionally distinguishable in respect to diagnostic signals $s_j \in S$ iff there is a diagnostic signal for which the subsets of its values corresponding to these faults are disjoint.

$$f_k R_{UD} f_m \Leftrightarrow \exists_{s_j \in S} V_{jk} \cap V_{jm} = \emptyset . \tag{9}$$

Definition 4. Faults $f_k, f_m \in F$ are strongly distinguishable in respect to diagnostic signals $s_j \in S$ iff there are at least two diagnostic signals for which the subsets of values corresponding to these faults are disjoint.

The above definitions were formulated on the assumption that all fault symptoms have occurred. Therefore, they are concerned for the inference approaches based on the fault signatures.

4 Metrics of Fault Distinguishability

Quantitative methods for determining the distinguishability of a fault pairs are well known. In the simplest case, fault distinguishability can be determined by calculating Hamming distance of signatures of both faults.

A commonly used metrics is the so-called diagnosability degree [16, 20, 23, 25]. Its value is calculated for binary diagnostic signals in two steps during the design stage of a diagnostic system. Firstly, the set of all considered faults is divided into disjoint subsets of indistinguishable faults. These subsets are called as elementary blocks or D -classes. In a second step, the number of D -classes, denoted as D_c , is divided by the total number of all considered faults. The resulting ratio is called the diagnosis accuracy

$$D^{DEG} = \frac{D_c}{card(F)} . \tag{10}$$

If all faults are distinguishable, the number of D -classes is equal to the number of all faults and the diagnostic accuracy is equal to 1. The value of this measure is in the range $[0, 1]$. However, the diagnosis accuracy D^{DEG} does not apply for FIS , as it does not take into consideration case of conditional distinguishability.

A proposition for a metric of fault distinguishability in a FIS has been given in [19]. However, it is computationally very complex. In turn, a simple metric of fault distinguishability was defined in [9] as

$$D^{ACC} = \left[\frac{\sum_{d_i \in D} card(d_i)}{card(D)} \right]^{-1} , \tag{11}$$

where d_i denotes the i^{th} diagnosis in the set D of all diagnoses.

Another metric is the fault isolability index γ [21, 22]. It is defined as the number of distinguishable pairs of faults.

5 New Metrics of Fault Distinguishability

The scope of the practical utility of known metrics of fault distinguishability is usually limited to methods that use the same way of notion the relationship fault-diagnostic signal values. Therefore, we will propose such two new measures of fault distinguishability, denoted hereafter by the symbols Δ and Δ^* , which, in contrast to those known and characterized in Sect. 4, will allow for a more comprehensive, rather than just a piecemeal, account of the degree of fault distinguishability.

Specifically, the metrics proposed in this paper will enable fault distinguishability analysis for all diagnostic inference based on *FSM* and *FIS* with binary and trinary diagnostic signal values. This will allow an objectified evaluation of the performance of the considered diagnostic systems.

The definition of the metrics of distinguishability Δ considers and valorises different types of fault distinguishability. In particular, the definition takes into account indistinguishability, conditional distinguishability and weak or strong distinguishability. The essence and, at the same time, the advantage of this metrics is that it simplistically, and thus practically, introduces arbitrary weighting coefficients reflecting the preference for the desired type of fault distinguishability in the diagnostic system. The metrics of distinguishability Δ is applicable wherever the determination of the type of distinguishability is possible.

Definition 5. The fault distinguishability metric Δ of the diagnostic system be:

$$\Delta = \frac{\sum_{K=1}^K \sum_{m=k+1}^K \delta_{km}}{0.5K(K-1)}, \quad (12)$$

where: $\Delta \in (0, 1]$; δ_{km} - a weighting factor reflecting the preference for the type of distinguishability of the pair of faults $\langle f_k, f_m \rangle$ $k \neq m$;

$$\delta_{km} = \begin{cases} 0.0 & \text{if faults } f_k \text{ and } f_m \text{ are indistinguishable} \\ 0.5 & \text{if faults } f_k \text{ and } f_m \text{ are conditionally distinguishable} \\ 1.0 & \text{if faults } f_k \text{ and } f_m \text{ are weak or strong distinguishable.} \end{cases}$$

The definition of the metrics Δ identically treats weak and strong distinguishability. However, the strong distinguishability is an important and desirable for a diagnostic system, as it describes the system's robustness to momentary changes in the values of diagnostic signals caused, for example, by interference or measurement noise. It is therefore also desirable to distinguish and indicate separate preferences for weak and strong distinguishability. For this reason, we will propose a modified definition of the fault distinguishability Δ^* .

Definition 6. The fault distinguishability metrics Δ^* of the diagnostic system be:

$$\Delta^* = \frac{\sum_{K=1}^K \sum_{m=k+1}^K \delta_{km}^*}{0.5K(K-1)}, \quad (13)$$

where: $\Delta^* \in (0, 1]$; δ_{km}^* - a weighting factor reflecting the preference for the type of distinguishability of the pair of faults $\langle f_k, f_m \rangle$ $k \neq m$;

$$\delta_{km}^* = \begin{cases} 0.00 & \text{- if faults } f_k \text{ and } f_m \text{ are indistinguishable} \\ 0.25 & \text{- if faults } f_k \text{ and } f_m \text{ are conditionally distinguishable} \\ 0.50 & \text{- if faults } f_k \text{ and } f_m \text{ are weak distinguishable} \\ 1.00 & \text{- if faults } f_k \text{ and } f_m \text{ are strong distinguishable.} \end{cases}$$

6 Calculation Example

To illustrate the properties of the proposed fault distinguishability metrics, we will refer to the relatively simple example of the diagnostic system presented in [10, 12]. However, due to the limitations in the paper's volume, the calculation example will be limited solely to an analysis of the fault distinguishability in *FSM* and *FIS* excluding a case of diagnosis based on timed sequences of fault symptoms.

The system to be diagnosed is a classical assembly of two serially connected liquid tanks. The choice of this system was dictated not only by its simplicity but also by the transparency and commonly understanding the physics of the flow phenomena.

It was assumed that two types of faults are possible in both tanks: liquid leakage from tanks f_1 and f_2 and obliteration of pipelines f_3 and f_4 . It was also assumed that the parameters of the diagnosed system are known, including: the cross-sectional areas of both tanks: A_1 and A_2 , the nominal cross-sectional area S_1 of the pipeline connecting tanks, the nominal cross-sectional area S_2 of the outlet pipeline from the second tank, and the flow contraction coefficients α_1 and α_2 of the outflows from both tanks.

The two-tank system is equipped with instrumentation for measurement of the fluid inflow F entering the first tank, the fluid level L_1 in the first tank and the fluid level L_2 in the second tank. It was also assumed that there could be faults in the instrumentation paths. A collection of all considered faults is shown in Table 1. By analogy with [10, 12], it was assumed that phenomenological partial models would be used for fault detection. In these models, the influence of disturbances and measurement noise was neglected. The following three flow balance models were used for fault detection:

$$A_1 \frac{dL_1}{dt} - F + \alpha_1 \cdot S_1 \sqrt{2g(L_1 - L_2)} = r_1 \quad (14)$$

$$A_2 \frac{dL_2}{dt} - \alpha_1 \cdot S_1 \sqrt{2g(L_1 - L_2)} + \alpha_2 \cdot S_2 \sqrt{2g \cdot L_2} = r_2 \quad (15)$$

$$A_1 \frac{dL_1}{dt} + A_2 \frac{dL_2}{dt} - F + \alpha_2 \cdot S_2 \sqrt{2g \cdot L_2} = r_3. \quad (16)$$

Table 1. Faults considered in a two-tank system

Symbol	Description
f_1	Leakage from tank #1
f_2	Leakage from tank #2
f_3	Obliteration in the pipeline connecting both tanks
f_4	Obliteration in the outflow pipe from tank #2
f_5	Faulty measurement of liquid inflow rate F
f_6	Faulty measurement of liquid level L_1 in the tank #1
f_7	Faulty measurement of liquid level L_2 in the tank #2

Table 2. *FSM* for a two-tank system

S/F	f_1	f_2	f_3	f_4	f_5	f_6	f_7	V_j
s_1	1	0	1	0	1	1	1	0, 1
s_2	0	1	1	1	0	1	1	0, 1
s_3	1	1	0	1	1	1	1	0, 1

Table 3. *FIS* for a two-tank system

S/F	f_1	f_2	f_3	f_4	f_5	f_6	f_7	V_j
s_1	-1	0	+1	0	-1, +1	-1, +1	-1, +1	-1, 0, +1
s_2	0	-1	-1	+1	0	-1, +1	-1, +1	-1, 0, +1
s_3	-1	-1	0	+1	-1, +1	-1, +1	-1, +1	-1, 0, +1

Table 4. Comparison of values of fault distinguishability metrics obtained for *FSM* and *FIS*

Metrics	<i>FSM</i>	<i>FIS</i>	<i>FSM</i>	<i>FIS</i>
Number of tests	2	2	3	3
D^{DEG}	0.429	–	0.571	–
D^{ACC}	0.429	0.571	0.571	0.714
Δ	0.762	0.881	0.857	0.928
Δ^*	0.476	0.583	0.619	0.678
$I - C - W - S^a)$	5-0-12-4	1-3-11-6	3-0-10-8	1-1-10-9

a) Number of pairs of faults: I – indistinguishable, C – conditionally distinguishable, W – weakly distinguishable, S – strongly distinguishable

Based on these models, the residuals r_1 , r_2 and r_3 were generated. The results of their evaluation in the form of binary and trinary values of the diagnostic signals s_1 , s_2 and s_3 are presented in Tables 2–3. Both were created with the usage of expert knowledge of the fault-diagnostic signal relationship.

Two different subsets of diagnostic tests will be considered in the example. The first subset contains tests based on all three models (14–16), while the second uses only the first two models (14–15). The calculated values of the fault distinguishability metrics for both subsets are shown in Table 4.

It should be noted that there is little point in comparing the values of distinguishability of different metrics. Each metric is used individually at the design stage to compare and evaluate different solutions of a diagnostic system. Unfortunately, due to the limited volume of this paper, it is not possible to present the full cycle of diagnostic system design. However, the four cases analysed in this section allow the following observations to be made:

- a) The fault distinguishability metric D^{DEG} , unlike the others, is unsuitable for use with trinary diagnostic signals;
- b) All metrics considered show higher values for a diagnostic system based on three tests compared to a system based on two tests. This shows the possibility to use these metrics for selection of the optimal set of tests by a given set of measurements.
- c) The calculation of the D^{ACC} requires the determination of all possible diagnoses and depends on the type of inference method used (based on signatures or symptoms).
- d) The Δ^* metrics is sensitive to strong fault distinguishability. It should be noted that strong distinguishability protects against false diagnoses caused by single false result of a diagnostic test.

7 Conclusions

This paper proposes two new simple and practically useful metrics of fault distinguishability. An important advantage of introducing both metrics is that they can be used to assess and compare the quality of fault isolation for a wide class of automated diagnostic systems based on binary and multivalued diagnostic signals, and optionally on elementary timed sequences of fault symptoms.

These metrics are mainly applicable at the design stage of a diagnostic system. This comes from the fact that they define a criterium useful for optimal or suboptimal selection of measurements and diagnostic tests.

In addition, the introduced metrics are characterized by a certain degree of application flexibility. This is because they provide the possibility of shaping the values of weighting coefficients, and thus reflecting individualized preferences of the type of fault distinguishability.

References

1. Blanke, M., Kinnaert, M., Lunze, J., Staroswiecki, M.: Diagnosis and Fault-tolerant Control. Springer, Heidelberg (2015). <https://doi.org/10.1007/978-3-662-47943-8>
2. Chen, J., Patton, R.: Robust Model Based Fault Diagnosis for Dynamic Systems. Kluwer Academic Publishers, Boston (1999)

3. Cordier, M.O., Dague, P., Lévy, F., Montmain, J., Staroswiecki, M., Travé-Massuyés, L.: Conflicts versus analytical redundancy relations: a comparative analysis of the model based diagnosis approach from the artificial intelligence and automatic control perspectives. *IEEE Trans. Syst. Man Cybernet. B: Cybernet.* **34**(5), 2163–2177 (2004)
4. Gertler, J.: *Fault Detection and Diagnosis in Engineering Systems*. Marcel Dekker Inc., (1998)
5. Gertler, J., Singer, D.: A new structural framework for parity equation based failure detection and isolation. *Automatica* **26**(2), 381–388 (1990)
6. Huang, Y., Gertler, J., McAvoy, T.J.: Sensor and actuator fault isolation by structured partial PCA with nonlinear extensions. *J. Process Control* **10**(5), 459–469 (2000)
7. Korbicz, J., Kościelny, J.M., Kowalczyk, Z., Cholewa, W.: *Fault Diagnosis. Models, Artificial Intelligence, Applications*. Springer, Berlin (2004). <https://doi.org/10.1007/978-3-642-18615-8>
8. Kościelny, J.M.: Recognition of fault in the diagnosing process. *Appl. Math. Comput. Sci.* **3**(3), 559–572 (1993)
9. Kościelny, J.M.: *Diagnostyka zautomatyzowanych procesów przemysłowych*. Akademicka Oficyna Wydawnicza EXIT, Warszawa (2001)
10. Kościelny, J.M., Bartyś, M.: A new method of diagnostic row reasoning based on trivalent residuals. *Expert Syst. Appl.* **214**(119116) (2023)
11. Kościelny, J.M., Bartyś, M., Syfert, M., Szyber, A.: A graph theory-based approach to the description of the process and the diagnostic system. *Int. J. Appl. Math. Comput. Sci.* **32**(2), 213–227 (2022)
12. Kościelny, J.M., Bartyś, M., Szyber, A.: Diagnosing with a hybrid fuzzy-Bayesian inference approach. *Eng. Appl. Artif. Intell.* **104**, 1–11 (2021)
13. Kościelny, J.M., Syfert, M., Rostek, K., Szyber, A.: Fault isolability with different forms of faults-symptoms relation. *Int. J. Appl. Math. Comput. Sci.* **26**(4), 815–826 (2016)
14. Kościelny, J.M., Syfert, M., Wnuk, P.: Diagnostic row reasoning method based on multiple-valued evaluation of residuals and elementary symptoms sequence. *Energies*, **14**(2476), (2021)
15. Kościelny, J.M., Syfert, M., Wnuk, P.: Diagnostic column reasoning based on multiple-valued evaluation of residuals and elementary symptoms sequence. *Energies* **15**(2614), (2022)
16. Krysander, M., Åslund, J., Nyberg, M.: An efficient algorithm for finding minimal overconstrained subsystems for model-based diagnosis. *IEEE Trans. Syst. Man Cybernet. Part A: Syst. Hum.* **38**(1), 197–206 (2007)
17. Patton, R., Frank, P., Clark, R. (eds.): *Issues of Fault Diagnosis for Dynamic Systems*. Springer-Verlag, Berlin, Heidelberg, New York (2000)
18. Puig, V., Schmid, F., Quevedo, J., Pulido, B.: A new fault diagnosis algorithm that improves the integration of fault detection and isolation. In: *44th IEEE Conference on Decision and Control*, pp. 3809–3814 (2005)
19. Rostek, K.: *Generalized metric of fault distinguishability for diagnostics of industrial processes*. PhD thesis, Institute of Automatic Control and Robotics of Warsaw University of Technology (2018)
20. Rouissi, F., Hoblos, G.: Fault tolerance in wind turbine sensor systems for diagnosability properties guarantee. In: *Conference on Control and Fault-Tolerant Systems (SysTol)*, pp. 523–528. IEEE (2013)

21. Sarrate, R., Blesa, J., Nejari, F., Quevedo, J.: Sensor placement for leak detection and location in water distribution networks. *Water Sci. Technol.: Water Supply* **14**(5), 795–803 (2014)
22. Sarrate, R., Nejari, F., Rosich, A.: Sensor placement for fault diagnosis performance maximization under budgetary constraints. In: *Proceedings of the 2nd International Conference on Systems and Control*, pp. 178–183 (2012)
23. Spanache, S., Escobet, T., Travé-Massuyés, L.: Sensor placement optimisation using genetic algorithms. In: *Proceedings of the 15th International Workshop on Principles of Diagnosis (DX-04)*, pp. 179–184 (2004)
24. Travé-Massuyés, L.: Bridges between diagnosis theories from control and AI perspectives. In: Korbicz, J., Kowal, M. (eds.) *Intelligent Systems in Technical and Medical Diagnostics*. volume 230, pp. 441–452. Springer, Heidelberg (2014)
25. Yassine, A., Ploix, S., Flaus, J.M.: A method for sensor placement taking into account diagnosability criteria. *Int. J. Appl. Math. Comput. Sci.* **18**(4), 497–512 (2008)



Active Noise Control with Passive Error Signal Shaping - A Critical Case Study

Małgorzata I. Michalczyk^(✉) 

Silesian University of Technology, 44-100 Gliwice, Poland
malgorzata.michalczyk@polsl.pl

Abstract. In the paper an application of passive residual error signal shaping in active noise control (ANC) system is revisited. The shaping band-pass filter was designed to attenuate some frequency range of the disturbance (denoted by the shaping filter pass-band) and exclude some frequency range of the disturbance (denoted by the shaping filter stop-band) from operation of the internal model ANC system using the filtered-x least mean squares algorithm. The simulations results show that depending on the time horizon the algorithm may attenuate the disturbance in the whole frequency range. Thus, it can be impossible to exclude a frequency range from the adaptive ANC system operation by means of the shaping filter. This means that passive residual error signal shaping in adaptive ANC systems does not work as expected.

Keywords: Active noise control · FxLMS · Residual noise shaping

1 Introduction

Active noise control (ANC) systems are used to attenuate unwanted noise [11]. In general, the goal of ANC systems is to attenuate the noise as much as possible. At best, the noise can be cancelled, so that the ANC system user can hear residual error signal in the form of a white noise, which can be strange and inconvenient. To deal with this problem, the residual error signal can be shaped to be perceived as more natural. It can be done in two ways: passively or actively. In a passive way the existing residual error signal is shaped by means of a specially designed filter. To shape residual error signal actively, an additionally generated signal, which can be further shaped, should be added to the system. As such signal a specially generated random signal should be used [5–7].

In adaptive ANC systems with filtered-x least mean squares (FxLMS) algorithm a passive method of residual error signal shaping was proposed in [12]. It uses an additional *a priori* designed shaping filter to shape spectral properties of the residual error signal. Based on this algorithm, so called psychoacoustic ANC systems are built to improve unwanted noise attenuation in terms of hearing perception [1] applying a special filter, e.g., an A-weighting filter. The residual error shaping band-stop or band-pass filter could allow also for excluding some range of frequencies from noise attenuation to retain a signal carrying valuable audible information, like an engine noise [18], speech or an alarm signal.

In [16], a behavior of the residual error shaping algorithm based on FxLMS algorithm was investigated in simulation experiments, with simplified and real-world plant models modeling an adaptive ANC system creating spatial zones of quiet in a reverberant enclosure. The algorithm was applied in the feedforward ANC system structure using a reference microphone to acquire a reference signal, carrying the reference information about the noise to be attenuated. The residual error signal frequency properties were successfully shaped; however, it was shown, that the assumptions made in [12] were not met and the nonlinearity of the adaptation feedback present in adaptive ANC systems should not be neglected.

The aim of the paper is to take a closer look at how the FxLMS algorithm with the passive residual error signal shaping works. The FxLMS algorithm with the passive residual error signal shaping is applied in the feedback ANC system. Commonly, a feedforward ANC system structure is applied to take advantage of the difference of the sound wave propagation times: the time required for the sound wave to go through the disturbance path and the time required for processing of the reference signal through the control and secondary paths. If the disturbance's travel lasts longer than the signal processing, the disturbance can be attenuated thoroughly. In the other case only the periodic disturbance (easy to predict) can be cancelled, but the random disturbance can be attenuated partially only [4, 8, 13]. In [10] it was shown how the ANC system efficiency decreases along with the increase of the time of signal processing. In particular it was shown, that the narrower the noise band is (the more correlated the disturbance), the more efficient the ANC system is. This rule applies also to the feedback ANC systems, which are utilized when there are problems with the reference signal acquisition. When the unwanted noise source cannot be located, there are many uncorrelated unwanted noise sources, or the signal processing routine delays control signal too much to obtain time precedence to implement feedforward control, the reference signal can be predicted using the Internal Model Control (IMC, [17]) structure [3, 11]. Economic factors may decide on the choice of internal model ANC system structure as well.

The paper is organized as follows: (1) the internal model active noise control system is described, (2) the results of two simulation experiments are presented and (3) discussed.

2 Internal Model Active Noise Control System

2.1 Idea of an Adaptive Active Noise Control

In order to attenuate unwanted noise, most often feedforward active noise control systems are applied. They utilize two signal sensors, for acquisition of a reference signal and an error signal. The reference signal is a kind of disturbance signal estimate, carrying information about the disturbance signal values (when obtained with precedence or if the disturbance is periodic) or its properties (otherwise). In the feedforward ANC system with an acoustic reference sensor, the reference signal is to be estimated due to the acoustical feedback between the control loudspeaker and the reference microphone.

The similar problem is solved in the internal model ANC system structure (block diagram shown in Fig. 1 (left)), when the direct acquisition of the reference or disturbance signal is impossible. Then, the disturbance signal $d(i)$ is picked up by an error microphone along with the control sound coming from the control loudspeaker, resulting in the error

signal $e(i)$. The disturbance signal $\hat{d}(i)$ is estimated by subtraction of the control signal $u(i)$, filtered by the model of the secondary path $\hat{S}(z^{-1})$, from the error signal $e(i)$, picked up by the error microphone. Then the $\hat{d}(i)$ signal is processed by the controller filter $W(z^{-1})$ giving the control signal $u(i)$, that drives the control loudspeaker. The control loudspeaker, an acoustic space, the error signal and the electronic part constitute the secondary path $S(z^{-1})$. The disturbance estimate is filtered through the model $\hat{S}(z^{-1})$ and the resulting signal along with the error signal are used to adjust controller filter $W(z^{-1})$ weights by the FxLMS adaptive algorithm (the most popularly applied in ANC [2, 8, 11]), to obtain minimum variance of the error signal $e(i)$.

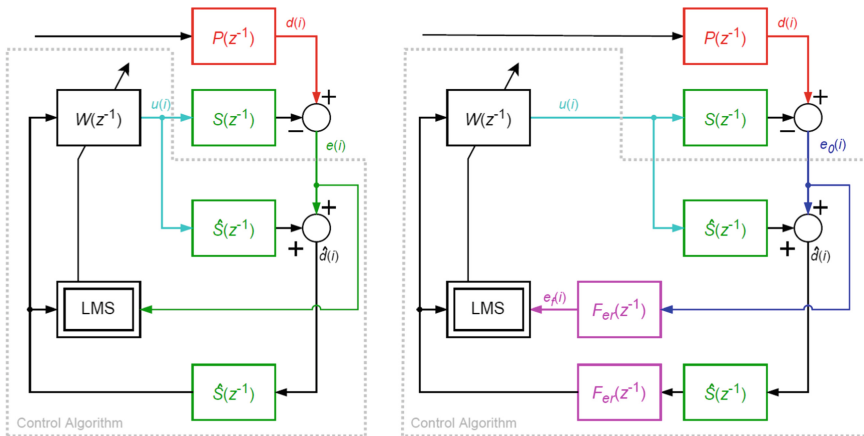


Fig.1. The block diagrams of the internal model ANC systems using the FxLMS algorithm (left) and the FxLMS algorithm with the residual error shaping filter (right).

2.2 Passive Error Signal Shaping

If the spectral properties of the error signal should meet some expectations, an adaptive control algorithm with the passive residual error shaping filter may be applied. In [12] an application of additional filter shaping the spectral properties of the residual error signal for the use with the FxLMS control algorithm was proposed. The residual error shaping filter $F_{er}(z^{-1})$ is placed on the error signal way, consequently adaptation LMS routine is fed with the filtered error signal $e_f(i)$ (Fig. 1(right)). In order to assure the ANC system convergence, the shaping filter is also placed along with the secondary path model in the path of the disturbance signal estimate fed to the adaptation routine. In such an ANC system, two error signals can be distinguished: the filtered error signal $e_f(i)$ used for adaptation of controller filter weights and the observed error signal $e_0(i)$ picked up by the error microphone, corresponding to what the ANC system user can hear.

The authors of the discussed algorithm assumed in [12] that $e_f(i) = F_{er}(z^{-1}) e_0(i)$. It might be true, if the nonlinear adaptation feedback was neglected. However, this assumption was experimentally proved to be wrong [16]. In [16] the behavior of the FxLMS algorithm with the residual error signal shaping filter was observed in simulation experiments, conducted with simplified and real-world plant models, modeling the feedforward ANC system creating spatial zones of quiet in a reverberant enclosure. Hereby the problem of residual error signal shaping in the internal model structure of the ANC system is revisited.

3 Simulation Experiments

To observe the performance of the internal model ANC system with the passive residual error shaping filter in comparison to the classical FxLMS algorithm a number of simulation experiments were conducted. The ANC system, that was modeled, reflected some properties of the real-world ANC system creating spatial local zones of quiet in an enclosure of 70 m³ cubature operating with 500 Hz sampling frequency [15]. The frequency response magnitudes of the disturbance path $P_d(z^{-1})$ and the secondary path $S(z^{-1})$ models are shown in Fig. 2. Due to dynamical complexity of the electro-acoustic plant in the enclosure, these magnitudes are very rugged. This is the reason for parameterization problems when applying FxLMS based control algorithm for ANC [2, 11, 14]. To focus on adaptive algorithm properties, the simplified ANC system paths models were used. The differences in magnitude inside the pass-band were neglected. Filters with similar spectral properties, but with different time delays, were used both for the disturbance and the secondary paths modeling. The band-pass FIR filters with almost linear phase characteristic were used with the bandwidth between 30 Hz and 180 Hz and the high attenuation in the stop-band (Fig. 2), corresponding to the real plant models. The time delays were set correspondingly 3 samples for the disturbance path and 5 samples for the secondary path, which corresponds with the real delay time values in ANC systems creating spatial local zones of quiet in enclosures. In such a system, also in feedforward control structure, only disturbances being highly correlated Gaussian noise, or a periodic signal may be thoroughly attenuated. In the simulations the perfect secondary path modeling $\hat{S}(z^{-1}) = S(z^{-1})$ was also assumed.

The shaping filter $F_{er}(z^{-1})$ was chosen as a pass-band FIR filter, passing the frequencies between 50 Hz and 100 Hz (Fig. 2). To observe the performance of the adaptive control algorithm the disturbance signal consisted of two sines of frequency 65 Hz (in the pass-band) and 150 Hz (in the stop-band). The values of the frequency response magnitude of the shaping filter $F_{er}(z^{-1})$ for these frequencies are depicted in Fig. 2 with black 'x', and for the frequency 150 Hz is equal -24.85 dB. Thus, the ANC system operates in the full frequency band, i.e. from 0 Hz to the Nyquist frequency, which in this example equals 250 Hz. Then, two bands are introduced by the shaping filter: one is a band in which there is noise control ($F_{er}(z^{-1})$ filter pass-band), and the other, in which no noise attenuation is expected ($F_{er}(z^{-1})$ filter stop-band). The signal with these two components and the white noise of low variance ($\sigma^2 = 10^{-6}$) was filtered through the simplified disturbance path model $P_d(z^{-1})$ (Fig. 2) to obtain the disturbance signal $d(i)$.

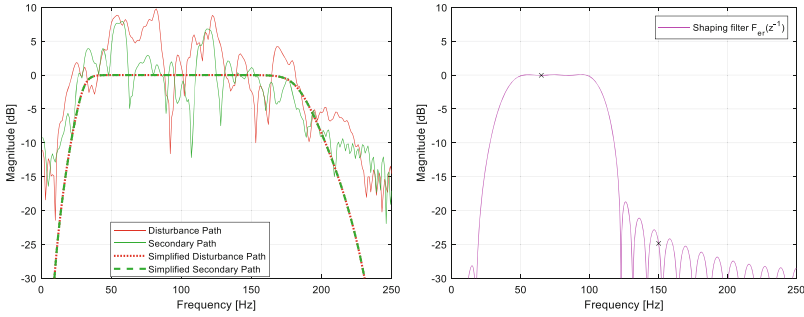


Fig. 2. Magnitudes of the frequency responses of the real-world and simplified plant models (left) and the magnitude of the frequency response of the band-pass residual error shaping filter (right).

The FIR controller filter $W(z^{-1})$ order and the step size μ value, assuring fast convergence and system stability, were chosen individually for each simulation experiment, the same for both algorithms. The controller filter with $N = 4$ coefficients was applied. It was a simple filter, however, sufficient for attenuating sum of two sines.

The signals analyzed during the simulation experiments were: disturbance signal $d(i)$, the same for both algorithms, error signal $e(i)$ for classical FxLMS algorithm, and for the FxLMS algorithm with the residual error shaping filter: filtered error signal $e_f(i)$ and the observed error signal $e_0(i)$, as well as the control signal $u(i)$. Estimated mean square values of the discrete-time signals were obtained by exponential smoothing (with the 0.999 factor, unless stated otherwise) of the squared signals. The signals power spectral densities were calculated by the averaging results obtained in non-overlapping windows over the $100 \cdot 5000$ samples period, after switching off the adaptation process (setting the step size μ value equal 0).

3.1 Experiment 1. What Happens to the Frequency Component in the Shaping Filter Stop-Band?

In the first simulation experiment a simple controller filter with $N = 4$ coefficients was applied to attenuate a periodic disturbance (sum of two sines) with added a low variance white noise ($\sigma^2 = 10^{-6}$). The simulation lasted 2'000 iterations (4 s). Mean square values of the discrete-time signals smoothed with the factor 0.9 are shown in Fig. 3 (left). Disturbance and error signals power spectral densities are in Fig. 3 (right).

The classical FxLMS algorithm attenuates the disturbance $d(i)$ to the lowest level, after about 1 s attenuation of 53 dB is obtained. The FxLMS algorithm with residual error shaping filter is faster at the beginning, during initial 100 algorithm iterations filtered error signal $e_f(i)$ reaches lower values. However, the attenuation of 48 dB is obtained after over 1.5 s. While the filtered error signal $e_f(i)$ is taken into calculations by the FxLMS algorithm with the residual error shaping filter, the signal that represents what the ANC system user can hear is the error signal $e_0(i)$ at the error microphone. This signal is amplified almost 1 dB. The analysis of the power spectral densities of the error signal shows what happened. The classical FxLMS algorithm attenuated both frequency components. The FxLMS algorithm with residual error shaping filter attenuated only

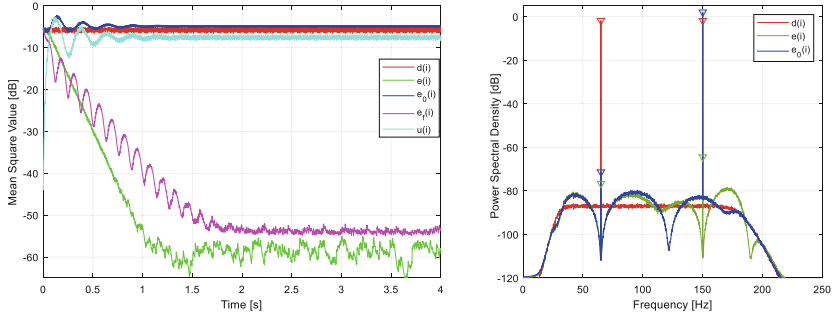


Fig. 3. (left) Estimated mean square values of the discrete-time signals: disturbance signal $d(i)$, error signal $e(i)$ for the FxLMS algorithm; error signals $e_0(i)$ and $e_f(i)$ and control signal $u(i)$ for the FxLMS algorithm with band-stop residual error shaping filter. (right) Estimated power spectral densities of the discrete-time signals: $d(i)$, $e(i)$ and $e_0(i)$ in Experiment 1.

65 Hz component (that in the shaping filter pass-band) and amplified by almost 4 dB 150 Hz component. Thus, the ANC system user does not hear the 65 Hz signal and can hear the 150 Hz signal as planned by selecting the spectral properties of the shaping filter $F_{er}(z^{-1})$.

After the signals are attenuated (both frequency components by classical FxLMS algorithm and the lower frequency component by the FxLMS algorithm with the residual error signal shaping filter), they do not change much. Situation seems stable for the next few seconds, the algorithms work as expected. The ANC systems are, however, designed to work over a much longer time horizon.

3.2 Experiment 2. What Happens in the Longer Time Horizon?

Simulation experiment 1. Was repeated in the longer time horizon of 50'000'000 iterations, i.e. 100'000 s (27 h 46 min 40 s). Again, the simple controller filter with $N = 4$ coefficients was applied to attenuate a periodic disturbance (sum of two sines) with added a low variance white noise ($\sigma^2 = 10^{-6}$). Mean square values of the discrete-time signals smoothed with the factor 0.999 are shown in Fig. 4 (left). Disturbance and error signals power spectral densities are in Fig. 4 (right).

In a longer time horizon, the error signal $e(i)$ remains at the same level, the classical FxLMS algorithm works without any changes. It turns out, however, that signals in the ANC system with the FxLMS algorithm with the residual error shaping filter are still changing for a long time. The filtered error signal $e_f(i)$ reaches lower mean square value level than the error signal $e(i)$ after more than 8 h, both control algorithms attenuate the disturbance thoroughly. After 20 min from the ANC system start, the error signal $e_0(i)$ mean square value level drops below the disturbance signal $d(i)$ mean square value level. It continues to drop steadily for the next 15 h, then being amplified again, until it finally stabilizes after 25 h. Additionally conducted simulations proved that all the signals attain stable levels over the next 27 h. It is worth emphasizing, that the control signal $u(i)$ mean square value stabilizes finally at the same level as a disturbance signal $d(i)$.

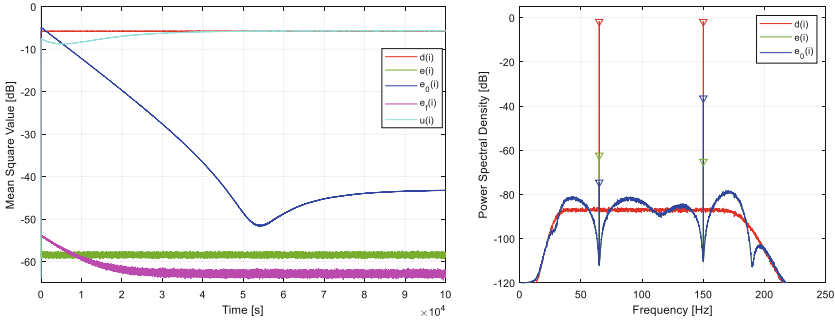


Fig. 4. (left) Estimated mean square values of the discrete-time signals: disturbance signal $d(i)$, error signal $e(i)$ for the FxLMS algorithm; error signals $e_0(i)$ and $e_f(i)$ and control signal $u(i)$ for the FxLMS algorithm with band-stop residual error shaping filter. (right) Estimated power spectral densities of the discrete-time signals: $d(i)$, $e(i)$ and $e_0(i)$ in Experiment 2.

Unfortunately, the results of Experiment 2 mean that the FxLMS algorithm with the residual error shaping filter attenuates the 150 Hz signal component, even if it is within the shaping filter $F_{er}(z^{-1})$ stop-band and should be passed through the system, according to the algorithm idea.

3.3 Discussion of the Simulation Results

The results of the simulations show that the FxLMS algorithm with the residual error shaping filter attenuates the disturbance in the entire frequency range. During the adaptation process, controller filter weights are constantly changed in such a way that the controller filter attenuates the frequencies present in the filtered error signal $e_f(i)$. Firstly, the dominating frequency components in the $F_{er}(z^{-1})$ filter pass-band are attenuated. Once the power of the dominating disturbance frequency component is sufficiently lowered, the adaptation algorithm starts to adjust controller filter weights to attenuate the other frequency components. The gain of the $F_{er}(z^{-1})$ filter in the stop-band is always nonzero, thus some frequency components from the stop-band of the $F_{er}(z^{-1})$ filter range are present in the filtered error signal, and the adaptive algorithm adjusts the controller filter $W(z^{-1})$ weights to attenuate them. Provided the ANC system operates long enough, it attenuates also the frequency components from the $F_{er}(z^{-1})$ filter stop-band, not fulfilling its role. Although a certain frequency range is intended to be passed through the system, it is ultimately a matter of time until it is attenuated. This situation may happen in any real-world ANC system, because such systems are designed to work in a time longer than a few seconds. Consequently, the algorithm specified in [12] does not allow for excluding any frequency range from the operation of the ANC system and thus is not suitable for shaping the residual error signal.

4 Conclusions

In the paper an application of passive residual error signal shaping in ANC system is revisited. The shaping band-pass filter was designed to attenuate some frequency range of the disturbance (denoted by the shaping filter pass-band) and exclude some frequency

range of the disturbance (denoted by the shaping filter stop-band) from operation of the internal model ANC system using the filtered-x least mean squares algorithm. The simulation results show that at a first glance the system seems to attenuate one disturbance frequency component and pass the second disturbance frequency component as intended. In a longer time horizon, however, the algorithm attenuates the disturbance in the entire frequency range. Although a certain frequency range is designed to be passed through the system, it is ultimately a matter of time until it is attenuated. Consequently, it can be impossible to exclude a frequency range from the adaptive ANC system operation by means of the shaping filter. This means that passive residual error signal shaping in adaptive ANC systems does not work as expected, what is implied by the nonlinear dynamic adaptation feedback.

The conclusion extends also to feedforward ANC systems, in which the disturbance signal is not estimated using the internal model but measured by the reference sensor.

Acknowledgement. The research reported in this paper has been supported by State Budget for Science, Poland: Silesian University of Technology grant number 02/050/BK_23/0032.

References

1. Bao, H., Panahi, I.M.S.: Using A-weighting for psychoacoustic active noise control. In: 31st Annual Int. Conf. of the IEEE EMBS (2009)
2. Elliott, S.J.: Signal Processing for Active Control, Academic Press (2001)
3. Elliott, S.J., Sutton, T.J.: Performance of feedforward and feedback systems for active noise control. *IEEE Trans. Speech Audio Process.* **4**(3), 214–223 (1996)
4. Eriksson, L.J., Allie, M.C., Greiner, R.A.: The selection and application of an IIR adaptive filter for use in active sound attenuation. *IEEE Trans. Acoustics Speech Signal Process.* **35**(4), pp. 433–437 (1987)
5. Figwer, J.: Adaptive Synthesis and Generation of Random Fields, Jacek Skalmierski Computer Studio, Gliwice (2008)
6. Figwer J., Synthesis and generation of random fields in nonlinear environment. In: Bartoszewicz, A., et al. (eds.) PCC 2020, AISC, vol. 1196. Springer (2020)
7. Figwer, J.: A New Approach to Acoustic Distributed Field Shaping. *Mechanics* **28**(3) (2009)
8. Hansen, C., Snyder, S.D.: Active Control of Noise and Vibration, Cambridge University Press (1997)
9. Kajikawa, Y., Gan, W.-S., Kuo, S.: Recent applications and challenges on active noise control. In: International Symposium on Image and Signal Processing and Analysis, ISPA, pp. 661–666 (2013). <https://doi.org/10.1109/ISPA.2013.6703821>
10. Kong, X., Kuo, S.M.: Study of Causality Constraint on Feedforward Active Noise Control Systems, *IEEE Trans. Circuits Syst. II: Analog Digital Signal Process.* **46**(2), 183–186 (1999)
11. Kuo, S.M., Morgan, D.R.: Active Noise Control Systems. J. Wiley & Sons Inc., New York, Algorithms and DSP Implementations (1996)
12. Kuo, S.M., Tsai, J.: Residual noise shaping technique for active noise control systems. *J. Acoustical Soc. Am.* **95**(3), 1665–1668 (1994)
13. Laugesen, S., Elliott, S.J.: Multichannel active control of random noise in a small reverberant room. *IEEE Trans. Speech Audio Process.* **1**(2), 241–249 (1993)
14. Michalczyk, M.I.: Parametrization of LMS-based control algorithms for local zones of quiet. *Archives Control Sci.* **15**(1), 5–34 (2005)

15. Michalczyk, M.I.: Comparison of feedforward and IMC controllers for active noise control system with moving error microphone: simulation results. In: 9th Conf. on Active Noise and Vibration Control Methods, Poland (2009)
16. Michalczyk M.I., Residual error shaping in active noise control - a case study. In: Bartoszewicz, A., et al. (eds.) PCC 2020, AISC, vol. 1196. Springer, Cham (2020). Doi: https://doi.org/10.1007/978-3-030-50936-1_60
17. Morari, M., Zafriou, E.: Robust Process Control. Prentice Hall, Englewood Cliffs, NJ (1989)
18. Ramos, P., Salinas, A., López, A., Masgrau, E.: Practical implementation of a multiple-channel FxLMS Active Noise Control system with shaping of the residual noise inside a van. In: Proc. of the 2002 Int. Symp. on Active Control of Sound (2002)

Fault Diagnosis and Fault-Tolerant Control



Neural Network Based Active Fault Diagnosis with a Statistical Test

Ivo Punčochár^(✉) and Ladislav Král

University of West Bohemia, Univerzitni 8, Plzen, Czech Republic
{ivop,ladkral}@ntis.zcu.cz

Abstract. The paper focuses on designing an active fault detector (AFD) for a non-linear stochastic system subject to abrupt faults. The neural network (NN) based models of the monitored system and their prediction error uncertainties are identified using historical input-output data obtained from the system under fault-free and all considered faulty conditions. The fault detector is based on a multiple hypothesis CUSUM-like statistical test that uses the identified NN models. The quality of decisions provided by such a detector is improved by a closed loop input signal generator. The input signal generator is represented by another NN and it is designed using a reinforcement learning method. The proposed AFD is illustrated by means of a numerical example.

1 Introduction

Fault detection has become an indispensable part of many systems in several application areas. The approaches to fault detection can be classified according to the interaction with a monitored system. While passive fault detection approaches only use the available data to make a decision about a fault in the monitored system [1, 6, 13], the active fault detection approaches introduce an auxiliary input signal that excites the monitored system in such a way that the faults are easier to detect [2, 10, 16, 17].

The active fault approaches use almost exclusively the multiple model framework, where fault-free mode and individual faulty modes of the monitored system are described by different models. The auxiliary input signal can be designed either to improve model discrimination or model change detection. In the model discrimination problem, the detector and the corresponding auxiliary input signal generator are designed under the assumption that there is no change of the mode during the time period for which the decision and the auxiliary input signal are designed [17]. On the other hand, in the model change detection problem, the possibility of mode change is taken into account when the AFD is designed [15]. Another classification criterion focuses on the assumption about the noises in the monitored system.

- Deterministic – the noises are assumed to be bounded signals and the auxiliary input signal can be designed to achieve guaranteed fault diagnosis [16],

- Stochastic – the noises are assumed to be stochastic processes with known properties and probability of incorrect decision is always nonzero [7, 17],
- Hybrid – combines the deterministic and stochastic description with the aim of reducing the conservative nature of the deterministic approach [9].

In this paper, the stochastic description of the system is assumed. The AFD consists of a passive detector (PD) and an auxiliary input signal generator. If information about mode switching is available the risk minimization and Bayesian approach can be used to design the PD [15]. If no such information is available, or it is deemed not to be reliable, the statistical tests like sequential probability ratio test or cumulative sum test can be employed [7, 14, 17]. If there are more than one alternative hypothesis, the design of a PD becomes more involved [3, 4, 12] and the multiple hypothesis active fault detection was considered, e.g., in [17].

The goal of this paper is to consider a non-linear system for which neither models of individual modes nor a models of mode switching are available. Instead, the proposed approach assumes that records of historical input-output data that includes fault-free and faulty modes of the monitored system complemented with diagnostic information are available. Therefore, it is assumed that the considered faults do not result into a complete failure of the monitored system that would prevent it from performing its intended purpose. Rather, the considered faults only decrease some performance characteristic of the monitored system and they are eventually fixed with some delay. Further, it is taken into account that the historical records contain mostly the input-output data for the fault-free mode of the system. The passive fault detector is designed to implement a CUSUM-like statistical test to detect the change from the fault-free mode to a faulty mode. The auxiliary input signal generator is designed using the policy gradient with parameter exploration (PGPE) method. It uses input-output trajectories and diagnostic information obtained using Monte Carlo simulations.

The paper is organized as follows. The problem is formulated in Sect. 2. The AFD is designed in Sect. 3. The proposed approach is illustrated by means of a numerical example in Sect. 4 and a summary is provided in Sect. 5.

2 Problem Formulation

This section provides the description of a monitored system, structure of AFD, and design criterion. It is assumed that the monitored system can be described at each time step $k \in \mathcal{T} = \{0, 1, \dots\}$ as

$$\mathbf{y}_{k+1} = \mathbf{f}_{\mu_k}(\mathbf{y}_{k-\ell_y:k}, \mathbf{u}_{k-\ell_u:k}) + \mathbf{e}_k, \quad (1)$$

where $\mathbf{y}_k \in \mathbb{R}^{n_y}$ is an output, $\mathbf{u}_k \in \mathcal{U} \subseteq \mathbb{R}^{n_u}$ is an input, $\mu_k \in \mathcal{M} = \{1, 2, \dots, M\}$ is a mode index, and $\mathbf{e}_k \in \mathbb{R}^{n_y}$ is a second-order stochastic process with the zero mean and unknown correlation function¹. The unknown non-linear function

¹ Note that $\mathbf{y}_{i:j}$ denotes the sequence of \mathbf{y}_k from the time step i up to the time step j .

$\mathbf{f}_i : \mathbb{R}^{n_y(\ell_y+1)} \times \mathcal{U}^{\ell_u+1} \mapsto \mathbb{R}^{n_y}$ represents the behaviour of the monitored system when the i -th mode is in effect. The lags $\ell_y \in \mathbb{Z}$ and $\ell_u \in \mathbb{Z}$ are also unknown. The sequence of initial inputs $\mathbf{u}_{-\ell_u:0}$ and outputs $\mathbf{y}_{-\ell_y:0}$ is described by an unknown pdf $p(\mathbf{y}_{-\ell_y:0}, \mathbf{u}_{-\ell_u:0})$. For the purpose of the active fault diagnosis, it is assumed that the monitored system starts in fault-free mode. The system switches to one of the faulty modes at an unknown time step and remains there indefinitely. Therefore, the mode index $\mu : \mathcal{T} \mapsto \mathcal{M}$ is modelled as a stochastic process of the following form

$$\mu_k = \begin{cases} 1 & \text{for } k = 0, 1, \dots, k_f^* - 1 \\ j^* & \text{for } k = k_f^*, k_f^* + 1, \dots \end{cases}, \quad (2)$$

where $k_f^* \in \mathcal{T}$ is an unknown time step of fault occurrence and $j^* \in \mathcal{M}_f = \mathcal{M} \setminus \{1\}$ is an unknown index of the fault mode that occurs. It is assumed that k_f^* and j^* are random variables with an unknown joint probability

$$P(k_f^* = k, j^* = j), \quad k \in \mathcal{T}, \quad j \in \mathcal{M}_f. \quad (3)$$

The AFD consist of a PD and an auxiliary input signal generator. It is assumed that the PD can utilize all past input output data to generate the current decision. On the other hand, the auxiliary input signal generator will be approximated by a function approximator that can effectively work only with past input-output data over a windows of a fixed length. Thus, the structure of the AFD is assumed to be

$$\begin{bmatrix} d_k \\ \mathbf{u}_k \end{bmatrix} = \rho_k(\mathbf{z}_{k,k,k}) = \begin{bmatrix} \sigma_k(\mathbf{z}_{k,k,k}) \\ \gamma(\mathbf{z}_{k,\ell'_y,\ell'_u}) \end{bmatrix}, \quad (4)$$

where $d_k \in \mathcal{M}$ is a decision, $\mathbf{z}_{k,i,j} = [\mathbf{y}_{k-i:k}, \mathbf{u}_{k-j:k-1}] \in \mathcal{Z}_{i,j} = \mathbb{R}^{n_y(i+1)} \times \mathcal{U}^j$ denotes the sequences of the past inputs and outputs with lag i and j , $\sigma_k : \mathcal{Z}_{k,k,k} \mapsto \mathcal{M}$ is a passive detector, and $\gamma : \mathcal{Z}_{k,\ell'_y,\ell'_u} \mapsto \mathcal{U}$ is an auxiliary input signal generator.

The lags $\ell'_y \in \mathbb{Z}$ and $\ell'_u \in \mathbb{Z}$ are selected by the designer and they are not directly related to the lags ℓ_y and ℓ_u of the system. The PD is assumed to be designed using a sequential statistical hypothesis test. Although one particular test will be presented in the sequel, we can also assume that the PD has been already designed and the only aim is to design the auxiliary input signal generator.

Finally, the auxiliary input signal generator γ is designed such that the following criterion is minimized

$$J(\gamma) = \lim_{F \rightarrow \infty} \mathbb{E} \left\{ \sum_{k=0}^F \lambda^k L(d_k, \mu_k) \right\}, \quad (5)$$

where $\lambda \in (0, 1)$ is a discount factor, $L : \mathcal{M} \times \mathcal{M} \mapsto \mathbb{R}^{++}$ is a detection cost function that satisfies $L(i, i) < L(i, j)$ for all $j \neq i$, and $\mathbb{E}\{\cdot\}$ is the expectation operator over all involved random variables. The discount factor λ and the detection cost function L are selected by a designer to meet desired design goals. A more detailed discussion regarding the discount factor and the detection cost function can be found, e.g., in [15].

3 Active Fault Detector Design

First, one-step ahead output predictors for the fault-free and all faulty modes are obtained from historical records of input-output data. They are subsequently used to design a PD. The auxiliary input signal generator is then designed using the PGPE algorithm. Besides these output predictors, the PGPE algorithm also requires a model of the stochastic process μ to perform Monte Carlo simulations of the predictors and designed PD.

3.1 System Identification

As mentioned in the introduction, it is assumed that historical records of input-output data complemented with diagnostic information are available. They are used to determine one-step ahead output predictors for the fault-free mode and all faulty modes of the monitored system and a stochastic model of μ . It is assumed that diagnostic information about the mode of the monitored system at each time step within the historical records has been obtained by means of an on-line diagnosis, human decision, post-processing diagnosis technique or their combination.

The input-output data are used to find one step ahead output predictor that is represented by a multi-layer perceptron (MLP) NN with a single hidden layer and hyperbolic tangent activation functions [11]. The one-step ahead prediction of the output \mathbf{y}_{k+1} for the i -th model is thus given as

$$\hat{\mathbf{y}}_{k+1}^i = \hat{\mathbf{f}}^i(\mathbf{x}_k, \boldsymbol{\theta}^i) = \boldsymbol{\theta}_1^i \tanh(\boldsymbol{\theta}_2^i \mathbf{x}_k) + \boldsymbol{\theta}_3^i, \quad (6)$$

where $\mathbf{x}_k = \left[\mathbf{y}_{k-\hat{\ell}_y:k}^T \quad \mathbf{u}_{k-\hat{\ell}_u:k}^T \quad 1 \right]^T \in \mathbb{R}^{n_x+1}$, $\boldsymbol{\theta}_1^i \in \mathbb{R}^{n_y \times m^i}$, $\boldsymbol{\theta}_2^i \in \mathbb{R}^{m^i \times n_x+1}$ and $\boldsymbol{\theta}_3^i \in \mathbb{R}^{n_y}$ are weights of the MLP NN that can be stacked column-wise into a single vector $\boldsymbol{\theta}^i \in \mathbb{R}^{m^i(n_x+n_y+1)+n_y}$, $m^i \in \mathbb{Z}$ is the number of neurons in the hidden layer. The lag $\hat{\ell}_y$ and $\hat{\ell}_u$ should reflect the values of true lags ℓ_y and ℓ_u . They can be optimized as part of NN training [5]. A dataset \mathcal{D}^i of input-output data that are used for training the i -th NN output predictor is retrieved from the historical records using the diagnostic information that tags each time step of the historical records with the correct index of the mode. Thus, for the time step k where mode i was deemed valid, the corresponding output \mathbf{y}_{k+1} and past input-output data $\mathbf{y}_{k-\hat{\ell}_y:k}$ and $\mathbf{u}_{k-\hat{\ell}_u:k}$ are included into the dataset \mathcal{D}^i . The resulting dataset is then used to find the weights $\boldsymbol{\theta}^i$ minimizing the one-step ahead prediction error. The training of the MLP NN is assumed to be performed using a well established algorithm that splits the dataset \mathcal{D}^i into training, validation, and test datasets. The test dataset is also used to estimate the covariance matrix of the prediction error $\hat{\boldsymbol{\Sigma}}_y^i$ that is subsequently used in the AFD design.

The description of the switching signal μ is obtained by writing the joint probability using the chain rule as

$$P(k_f^* = k, j^* = j) = P(k_f^* = k | j^* = j) P(j^* = j), \quad (7)$$

where $P(j^* = j)$, $j \in \mathcal{M}_f$ is an unknown probability distribution of fault occurrence and $P(k_f^* = k | j^* = j)$, $k \in \mathcal{T}$ is an unknown conditional probability of fault time occurrence. We assume that the time of fault occurrence k_f^* conditioned by the faulty model j^* has the geometric distribution

$$P(k_f^* = k | j^* = j) = p_j^k (1 - p_j), \quad (8)$$

where $k \in \mathcal{T}$ is an auxiliary variable and p_j is an unknown probability being in the fault-free mode. The probability p_j can be estimated from historical records using the maximum likelihood method. We isolate sequences of decisions that start in the fault-free mode after being at a faulty mode and ends up by faulty mode j . We assume that these realizations are independent and the probability p_j is estimated in the maximum likelihood sense as

$$\hat{p}_j = \arg \max_{p_j \in [0,1]} \prod_{i=1}^{N_I} p_j^{k_i} (1 - p_j), \quad (9)$$

where N_I is the number of sequences that were retrieved from the historical records and k_i is the number of steps before the switch to a faulty mode has occurred in the i -th sequence. The probability $P(j^* = j)$ for $j \in \mathcal{M}_f$ is estimated as relative occurrence of switching from fault-free model to the fault model j within the historical records.

3.2 Passive Detector Design

It is assumed that the PD is designed using a multiple hypothesis sequential statistical test. Assuming that k denotes the current time step, the hypothesis are

- \mathcal{H}_1 - the input-output samples \mathbf{y}_t and \mathbf{u}_t are generated by the system in the fault-free mode for all time steps $t = 0, 1, \dots, k$,
- \mathcal{H}_i - the input-output samples \mathbf{y}_t and \mathbf{u}_t are generated by the system in the fault-free mode for the time steps $t = 0, \dots, k_f - 1$ and by the system in the faulty mode i for time steps $t = k_f, \dots, k$.

The likelihood of $\mathbf{u}_{0:k}$ and $\mathbf{y}_{0:k}$ under the hypothesis \mathcal{H}_j is given as

$$p(\mathbf{y}_{0:k}, \mathbf{u}_{0:k}; \mathcal{H}_j) = \prod_{t=0}^{k_f-1} p(\mathbf{u}_t | \mathbf{y}_{0:t}, \mathbf{u}_{0:t-1}) p^1(\mathbf{y}_t | \mathbf{y}_{0:t-1}, \mathbf{u}_{0:t-1}) \prod_{t=k_f}^k p(\mathbf{u}_t | \mathbf{y}_{0:t}, \mathbf{u}_{0:t-1}) p^j(\mathbf{y}_t | \mathbf{y}_{0:t-1}, \mathbf{u}_{0:t-1}). \quad (10)$$

Since the auxiliary input signal generator does not depend on the mode of the monitored system and the terms up to the time step $k_f - 1$ are the same, the log likelihood ratio the hypotheses \mathcal{H}_i and \mathcal{H}_j can be written as

$$S_k^{ij}(k_f) = \ln \left(\frac{p(\mathbf{y}_{0:k}, \mathbf{u}_{0:k}; \mathcal{H}_i)}{p(\mathbf{y}_{0:k}, \mathbf{u}_{0:k}; \mathcal{H}_j)} \right) = \ln \left(\frac{\prod_{t=k_f}^k p^i(\mathbf{y}_t | \mathbf{y}_{0:t-1}, \mathbf{u}_{0:t-1})}{\prod_{t=k_f}^k p^j(\mathbf{y}_t | \mathbf{y}_{0:t-1}, \mathbf{u}_{0:t-1})} \right). \quad (11)$$

Although the prediction errors of the MLP NNs (6) does not have exact Gaussian distribution and can be correlated in time, it seems reasonable to approximate $p^j(\mathbf{y}_t | \mathbf{y}_{0:t-1}, \mathbf{u}_{0:t-1})$ as

$$p^j(\mathbf{y}_t | \mathbf{y}_{0:t-1}, \mathbf{u}_{0:t-1}) \approx \mathcal{N} \left\{ \mathbf{y}_t : \hat{\mathbf{y}}_t^j, \hat{\Sigma}_y^j \right\}, \tag{12}$$

where \mathcal{N} denotes the Gaussian pdf. The recursive version of the matrix CUSUM test proposed in [12] is employed as follows. The statistics of the matrix CUSUM test are computed as

$$S_k^{ij} = \max \left(0, S_{k-1}^{ij} \right) + s_k^{ij}, \tag{13}$$

where $i \in \mathcal{M}_f$ and for each i only $j \in \mathcal{M}_i = \mathcal{M} \setminus \{i\}$ are considered and the log likelihood ratio s_k^{ij} is computed as

$$s_k^{ij} = \ln \left(\frac{\mathcal{N} \left\{ \mathbf{y}_k : \hat{\mathbf{y}}_k^i, \Sigma_y^i \right\}}{\mathcal{N} \left\{ \mathbf{y}_k : \hat{\mathbf{y}}_k^j, \Sigma_y^j \right\}} \right). \tag{14}$$

The decision is generated as

$$d_k = \begin{cases} 1 & \text{if } \max_{i \in \mathcal{M}_f} \bar{S}_k^i \leq S_{\text{th}}, \\ \arg \max_{i \in \mathcal{M}_f} \bar{S}_k^i & \text{otherwise,} \end{cases} \tag{15}$$

where S_{th} is a selected threshold and \bar{S}_k^i is computed as $\bar{S}_k^i = \min_{j \in \mathcal{M}_i} S_k^{ij}$.

3.3 Auxiliary Input Signal Generator Design

The input signal generator is represented by another MLP NN. The parameters of this NN are found using the PGPE method that employs Monte Carlo simulations. The input signal generator is represented by the MLP NN

$$\mathbf{u}_k = \boldsymbol{\gamma} \left(\mathbf{z}_{k, \ell'_y, \ell'_u} \right) = \text{proj}_{\mathcal{U}} \left(\boldsymbol{\theta}_1 \tanh \left(\boldsymbol{\theta}_2 \bar{\mathbf{x}}_k \right) \right), \tag{16}$$

where $\text{proj}_{\mathcal{U}} : \mathbb{R}^{n_u} \mapsto \mathcal{U}$ is a projection operator that ensures that the input \mathbf{u}_k belongs to the set of admissible inputs \mathcal{U} . In principle, the proposed MLP NN (16) could use only the inputs and outputs of the monitored system, assuming that it has a sufficiently rich structure to provide a reasonable generator of auxiliary input signals. However, the practical experiments indicate that the convergence of the PGPE methods can be significantly improved if the predictions $\hat{\mathbf{y}}_k^i$ for $i = 1, \dots, M$ are also provided as inputs to the NN (16). Therefore, the vector $\bar{\mathbf{x}}_k$ is selected as

$$\bar{\mathbf{x}}_k^T = \left[\mathbf{y}_{k-\bar{\ell}_y:k} \ \mathbf{u}_{k-\bar{\ell}_u:k-1} \ \hat{\mathbf{y}}_{k-\bar{\ell}_y:k-1}^1 \ \cdots \ \hat{\mathbf{y}}_{k-\bar{\ell}_y:k-1}^M \ 1 \right], \tag{17}$$

where $\bar{\ell}_y$, $\bar{\ell}_u$, and $\bar{\ell}_y$ are lags of the outputs, inputs, and predictions, respectively. They must be selected by a designer. The details NN training using the PGPE are omitted as they can be found, e.g., in [8].

4 Numerical Example

The proposed approach is illustrated using a numerical example that represents a mathematical pendulum. The discrete-time state-space model of the pendulum with the sampling period $T_s = 0.05$ s is

$$X_{k+1} = \begin{bmatrix} X_{1,k} + T_s X_{2,k} + w_{1,k} \\ -\frac{T_s g}{\ell_{\mu_k}} X_{1,k} + \left(1 - \frac{T_s \beta_{\mu_k}}{m \ell_{\mu_k}^2}\right) X_{2,k} + \frac{T_s}{m \ell_{\mu_k}^2} u_k + w_{2,k} \end{bmatrix}, \quad (18)$$

$$y_k = [1 \ 0] X_k + v_k, \quad (19)$$

where $X_{1,k}$ [rad] the angle of displacement from the zero downward position, $X_{2,k}$ [rad s⁻¹] is the corresponding angular velocity, $X_k = [X_{1,k}, X_{2,k}]^T$ is the continuous part of the state, u_k [Nm] is the torque applied at the joint, w_k and v_k are the state noise and the measurement noise with zero mean and covariance matrices $8 \cdot 10^{-4} \mathbf{I}_2$ and $1 \cdot 10^{-3}$, respectively. Parameters of the system are given by the the gravitational acceleration $g = 9.81$ m s⁻², pendulum mass $m = 2$ kg. In addition, varying parameters of the pendulum length ℓ_{μ_k} [m] and the friction coefficient β_{μ_k} [N m² s⁻¹] define the fault-free and faulty modes of the system, respectively, and the corresponding values are shown in Table 1. For the numerical example we assume that the probabilities $P(j^*)$ and the parameters of the geometric distributions $P(k_f^* | j^*)$ are known, where $p_1 = p_2 = 0.98$ and $P(j^* = 2) = P(j^* = 3) = 0.5$.

The input-output data of length 5000 were generated using a state space model (18)–(19) for one fault-free and each of two faulty modes, respectively. The data were divided into training, validation and test set at a ratio of 0.7:0.15:0.15. The NN structure was chosen identical for all models with $m^{1,2,3} = 4$ neurons in the hidden layer and lags $\hat{\ell}_y = \hat{\ell}_u = 2$. The individual NN models were obtained using the Levenberg-Marquardt backpropagation algorithm. The quality of the NN models for test datasets is illustrated in Fig. 1, where the shape of histograms justifies (12).

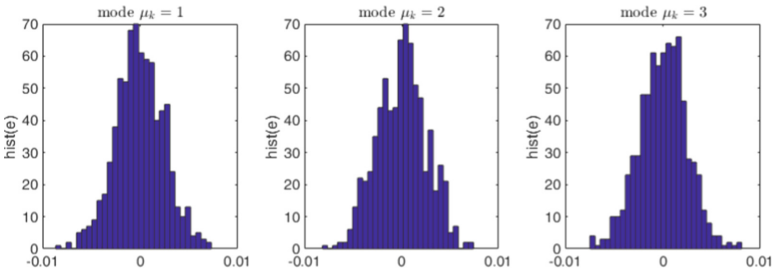


Fig. 1. A quality of the NNs models in a form of the prediction error histograms.

Using the NN models, the auxiliary input signal generator was designed using the another MLP NN with 20 neurons and lags $\bar{\ell}_y = 0$, $\bar{\ell}_u = 2$, and $\bar{\ell}_{\dot{y}} = 2$. The

NN parameters were optimized with the PGPE algorithm with following set-up: $\lambda = 0.99$, $\alpha = 1$, $\gamma_b = 0.1$, $N = 100$, $F_a = 360$, $\mathbf{m}_\theta = \mathbf{0}$, $\Sigma_\theta = 0.5\mathbf{I}$, where the parameter notation is taken from [8]. A typical trajectory of the system, input signal generator and detector is depicted in Fig. 2. The true mode μ_k and decision d_k (the middle plot) show the detector working correctly, with a delay in the decision averaging around 10 time instants. Similar detection quality was achieved for both considered faulty modes. The shape of the auxiliary input signal is close to the rectangular signal. This input signal excites the system to reveal its dynamical behavior for the given range of input signal $\mathcal{U} = [-10, 10]$. Another input signal, for example a random signal or a harmonic signal, results into a lower quality of fault detection.

Table 1. Parameters of the monitored system for fault-free and faulty modes

Mode	ℓ_{μ_k}	β_{μ_k}
Fault-free ($\mu_k=1$)	1	6
Fault 1 ($\mu_k=2$)	1	5
Fault 2 ($\mu_k=3$)	1.2	6

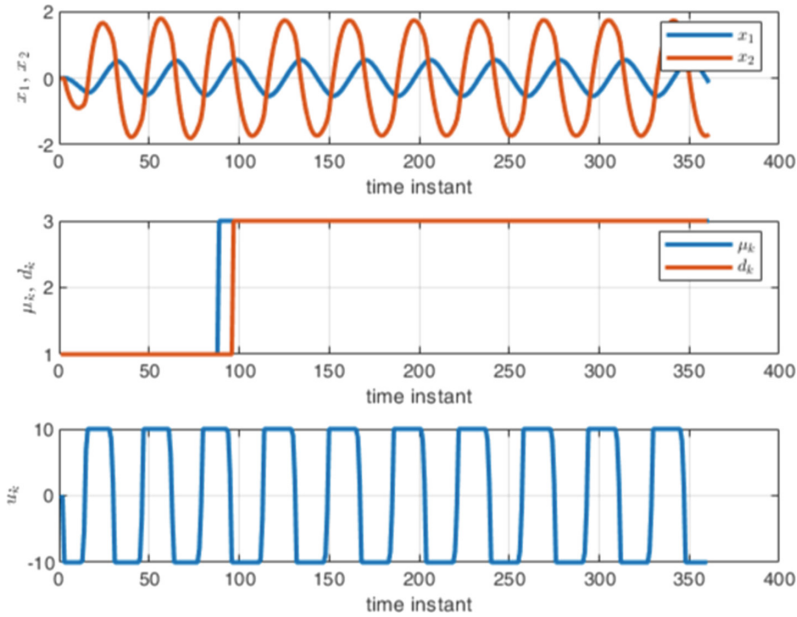


Fig. 2. Typical trajectories of the pendulum system using the proposed AFD for Fault 2 occurring in step 85 (top - displacement angle and angular velocity, middle - true system mode and decision generated by the detector, bottom - input signal produced by the auxiliary input signal generator).

5 Conclusion

The paper considers active fault detection for a unknown non-linear discrete-time stochastic system. The design of AFD is thus performed in two stages. First, one-step ahead predictors and model of switching between modes in obtained from the historical records of the input-output data. Second the passive fault detector based on a statistical test is designed and complemented with an auxiliary input signal generator that is designed using reinforcement learning technique. The proposed AFD design is successfully illustrated in a numerical example.

Acknowledgments. The work was supported by the Czech Science Foundation under grant 22-11101S.

References

1. Blanke, M., Kinnaert, M., Lunze, J., Staroswiecki, M.: *Diagnosis and Fault-Tolerant Control*. Springer, Heidelberg (2016). <https://doi.org/10.1007/978-3-662-47943-8>
2. Campbell, S.L., Nikoukhah, R.: *Auxiliary Signal Design for Failure Detection*. Princeton University Press, Princeton, NJ, USA (2004)
3. Golubev, G., Safarian, M.: A multiple hypothesis testing approach to detection changes in distribution. *Math. Methods Statist.* **28**(2), 155–167 (2019)
4. Haroutunian, E.A., Hakobyan, P.M.: Most powerful test for multiple hypotheses. In: *Proceedings of the 10th International Conference on Computer Science and Information Technology*, pp. 1–3. Yerevan, Armenia (2015)
5. Haykin, S.: *Neural networks: a comprehensive foundation*. Prentice Hall (1999)
6. Isermann, R.: *Fault-Diagnosis Applications*. Springer, Heidelberg, Germany (2011). <https://doi.org/10.1007/978-3-642-12767-0>
7. Kerestecioglu, F.: *Change Detection and Input Design in Dynamical Systems*. Research Studies Press, Taunton, England (1993)
8. Král, L., Punčochář, I.: Policy search for active fault diagnosis with partially observable state. *Adapt. Control Signal Process.* **36**(9), 2190–2216 (2022)
9. Marseglia, G.R., Scott, J.K., Magni, L., Braatz, R.D., Raimondo, D.M.: A hybrid stochastic-deterministic approach for active fault diagnosis using scenario optimization. In: *Proceedings of the 19th IFAC World Congress*, pp. 1102–1107. Cape Town, South Africa (2014)
10. Niemann, H.H.: A model-based approach to fault-tolerant control. *Int. J. Appl. Math. Comput. Sci.* **22**(1), 67–86 (2012)
11. Nørgaard, M., Ravn, O., Poulsen, N.K., Hansen, L.K.: *Neural Networks For Modelling And Control Of Dynamic Systems: A Practitioner's Handbook* (2000)
12. Oskiper, T., Poor, H.V.: Matrix CUSUM: a recursive multi-hypothesis change detection algorithm. In: *Proceedings of the 2001 IEEE International Symposium on Information Theory*, p. 19. Washington, DC, USA (2001)
13. Patton, R.J., Frank, P.M., Clark, R.N.: *Issues of Fault Diagnosis for Dynamic Systems*, 1st edn. Springer-Verlag, London, London, United Kingdom (2000). <https://doi.org/10.1007/978-1-4471-3644-6>
14. Poulsen, N.K., Niemann, H.H.: Active fault diagnosis based on stochastic tests. *Int. J. Appl. Math. Comput. Sci.* **18**(4), 487–496 (2008)

15. Punčochář, I., Šimandl, M.: On infinite horizon active fault diagnosis for a class of non-linear non-Gaussian systems. *Int. J. Appl. Math. Comput. Sci.* **24**(4), 795–807 (2014)
16. Raimondo, D.M., Marseglia, G.R., Braatz, R.D., Scott, J.K.: Closed-loop input design for guaranteed fault diagnosis using set-valued observers. *Automatica* **74**, 107–117 (2016)
17. Zhang, X.J.: *Auxiliary Signal Design in Fault Detection and Diagnosis*. Springer-Verlag, Berlin, Germany (1989). <https://doi.org/10.1007/BFb0009313>



Fault-Tolerant Fast Power Tracking Control for Wind Turbines Above Rated Wind Speed

Horst Schulte^(✉) and Nico Goldschmidt

Control Engineering Group, School of Engineering I,
University of Applied Sciences Berlin (HTW), Berlin, Germany
schulte@htw-berlin.de

Abstract. This paper proposes a model-based fault-tolerant control system for wind turbines with power tracking above the rated wind speed. Due to the increasing share of renewable energy sources such as wind turbines in power systems, the latter must provide fast power tracking for frequency control. So far, these power-tracking control systems are not yet equipped with fault-tolerance characteristics. Therefore, an fault-tolerant control design using a reconfiguration block is proposed. In the reconfiguration block, a proportional-multi-integral observer calculates actuator faults if such faults should occur and eliminates their influence on the rated power tracking controller. An analytical model-based design method for the reconfiguration block is presented in detail. System simulations are used to show how this will improve the reliability of future power-tracking wind turbines.

1 Introduction

Due to the increasing feed-in of renewable energies and the decreasing number of synchronous machines in conventional power plants, system services will have to be reliably provided by e.g. wind and photovoltaic plants in the future. A necessary auxiliary service is the provision of instantaneous reserves for frequency stability. For renewable energy sources connected to the grid via power converters to provide this service, i.e. PV power plants and wind energy systems must be equipped with active, fast power tracking of the primary converter system. The primary conversion includes converting the primary source, e.g., the irradiation or wind flow, into electrical power, which the secondary conversion process converts into suitable AC power for the grid.

Based on the results in [1], it is mandatory to consider the primary converter dynamics and a related control strategy to improve the power response by automatic control. To ensure stable and reliable operation even with an increasing share of renewable energy, it is necessary to enhance wind turbine control systems with power tracking functionality [2]. Wind turbines can then act as decentralized control systems in energy grids, continuously adjusting the currently generated power in response to the grid frequency [3]. Promising research results have been published in recent years to solve

this challenge. First, a heuristic¹ extension of the conventional PI controller with gain-scheduling control was proposed in [4, 5], and [6] to enable power tracking. Second, extending the operating range by dynamically controlling the generated power was considered in model-based approaches in [7–10]. Experimental validation of power tracking control systems through wind tunnel testing has been conducted in two studies to date: A gain-scheduled PI controller for a scaled wind turbine was presented in [11]. An LMI-based multiple-input and multiple-output pole-region controller design using $N_r = 252$ linearized local models numerically calculated by Taylor linearization from the NREL wind turbine model [12] was proposed and experimentally validated in [13].

The power tracking controllers proposed in previous studies cannot actively tolerate sensor and/or actuator faults. To close that gap, a controller structure and an analytical synthesis method are presented in this paper to extend a controller with power tracking capability in such a way that it shows fault-tolerant (FT) behavior. Based on the fault hiding approach [14, 15], a reconfiguration block is placed between the plant with actuator faults and the nominal power tracking controller. The reconfiguration block must meet the requirement that the behavior of the reconfigured system corresponds as closely as possible to the behavior of the nominal, i.e., fault-free system. Fault-tolerant control (FTC) of wind turbines without power tracking has been well-researched in recent years [16]. A study investigates also FTC concepts for wind turbines using the reconfiguration method with an observer concept but without power tracking should be mentioned here. In [17] it was shown that fault-tolerant control for wind turbines is feasible with the actuator fault scenarios investigated.

This work examines which limitations one has to accept and how robust the system has to be designed. The design is analytically based on an interpretable first-principle generic wind turbine model. Here, the reconfiguration is based on a continuous compensation of the controller output using estimated actuator faults. The estimation is done by a proportional-multi-integral observer using a polytopic representation of the wind turbine plant in Takagi-Sugeno form. The paper is organized as follows: First, the overall fault-tolerant power tracking control system is described in Sect. 2. In Sect. 3, the control-oriented state-space model for fault tolerant control (FTC) design is proposed. Section 4 describes in detail the synthesis of the reconfiguration block using fault reconstruction with a proportional-integral observer. Finally, in Sect. 5, a generic process model demonstrates the hypothesis of a fault-tolerant control for wind turbines using power tracking.

2 Fault Tolerant Power Tracking Control System

The fault tolerant power tracking control system consists of several components: The nominal open-loop power control, the nominal lower-level closed-loop speed control in combination with the reconfiguration block, and the fault estimator for the FTC characteristic. Further, as an interface to the grid frequency controller, a pre-processing verification unit of the higher-level power request $\Delta P_{g,req}$. The overall structure and the

¹ Heuristic in the sense that the controller is not model-based, is not reproducibly parameterized, and no closed-loop stability proof has been presented.

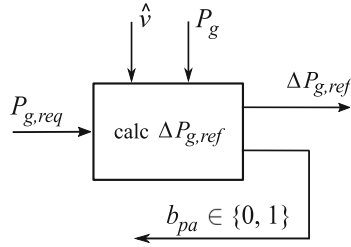


Fig. 2. Calculation of difference reference power $\Delta P_{g,ref}$ by (1) as a function of requested power $P_{g,req}$, available instantaneous power P_g and estimated effective wind speed \hat{v} with acknowledge bit b_{pa} for higher-level frequency grid controller

$b_{pa} = \{0, 1\}$, where “1” denotes that power is available. The function (1) with this additional feedback channel is illustrated as a system block in Fig. 2 that complements the power tracking control scheme in Fig. 1. If the maximum available power is below the rated power, i.e., $v < v_d$, the adjustable difference can no longer be calculated with $P_{g,rated}$. Instead, it must be determined from the moving average $P_{g,av}$ of the injected instantaneous power

$$P_{g,av}(t) = \frac{1}{N} \sum_{i=0}^{N-1} P_g(t - iT_s), \tag{2}$$

where denotes a constant sampling time T_s for $t = T_s i$ with $i = 1, \dots, N$.

2.2 Nominal Control Scheme Description

Conventional control systems for wind turbines operate them in the partial load range below the rated wind speed² and in full load above the rated wind speed. As long as the wind speed is below the rated wind speed, the controller tracks the optimum ratio of the rotor speed and wind speed by adjusting the generator torque T_g . Above the rated speed $v > v_d$, the wind turbine is operated at the rated speed of the rotor and, in established standard control concepts, with a constant generator-rated torque that cannot be adjusted. Instead, the rotor blade angle β is the manipulated variable with which the rotor torque is indirectly adjusted. For power tracking above the rated wind speed, the valid working range of the control systems must be extended to include T_g values below the rated value $T_{g,rated}$. The related control scheme is illustrated in Fig. 1. Assuming that the rotor speed ω_r is set to the fixed rated speed $\omega_{r,ref} = \omega_{r,rated} = \text{const}$, the power is tracked by varying the generator torque alone, with the rotor speed at the rated speed controlled by the rotor blade pitch angle. Another degree of freedom for power tracking is provided by variation of the rotor speed set point $\omega_{r,ref}$. Adjusting the rotor speed for fast power response is not suitable because

² The rated wind speed is also called design wind speed v_d .

the rotor's inertia strongly limits it. The power to be fed in is matched in the open loop by adjusting the generator torque ΔT_g , while the closed-loop speed control keeps the rotor speed at the rated set point.

From the view of automated control, the FTC functionality is achieved by an add-on system. For power tracking, a reconfiguration block is connected between the output of the nominal controller and the plant input $\mathbf{u} = (T_g, \beta_{ref})^T$. A fault estimator provides the reconfiguration block (see Fig. 1) with the estimated actuator fault, denoted $\hat{\mathbf{f}}_a$, at the signal level, which is reconstructed using the available measurement signals.

3 Control-Oriented Model for the Nominal Controller Synthesis

The control-oriented wind turbine model with the purpose of nominal power tracking design and fault estimation for FTC reconfiguration is given as follows

$$\begin{aligned} \dot{\mathbf{x}} &= \begin{pmatrix} f_1(\mathbf{x}, T_g, v) & f_2(\mathbf{x}, v) \\ 0 & -\frac{1}{\tau_\beta} \end{pmatrix} \mathbf{x} + \begin{pmatrix} 0 \\ \frac{1}{\tau_\beta} \end{pmatrix} u, \\ \mathbf{y} &= \begin{pmatrix} n_g & 0 \\ 0 & 1 \end{pmatrix} \mathbf{x}, \quad u = \beta_{ref} \end{aligned} \quad (3)$$

with the state vector $\mathbf{x} = (x_1, x_2)^T = (\omega_r, \beta)^T$, the reference value of the collective pitch angle β_{ref} as the system input and the functions

$$\begin{aligned} f_1(\mathbf{x}, T_g, v) &= \frac{1}{2} \left(\frac{\rho \pi R^3 v^2 c_{Q,1}(\mathbf{x}, v)}{J} - \frac{n_g T_g}{J} \right) \frac{1}{x_1}, \\ f_2(\mathbf{x}, v) &= \frac{\rho \pi R^3 v^2 c_{Q,2}(\mathbf{x}, v)}{2J} \frac{1}{x_2}, \end{aligned} \quad (4)$$

for $T_{g,min} \leq T_g \leq T_{g,max} = T_{g,rated}$, where T_g denotes generator torque, β the collective pitch angle of the blades, and ω_r denotes the rotor angular velocity. The upper limit of the generator torque is the rated power $T_{g,max} = T_{g,rated}$. The system parameters in (3) are ρ as the air density, R as the rotor radius, n_g as the gearbox ratio and J as the effective 1-DOF inertia determined by $J = J_r + n_g^2 J_g$ with the rotor inertia J_r and generator inertia J_g . To meet the control requirements for mechanical load and vibration mitigation by multi-criteria design, the underlying motion equation in (3) can be straightforwardly extended by the tower deflection, collective blade deflection, and torsional angle dynamics (2 to 4-DOF). However, the basic structure of the presented controller synthesis, in particular, considering the nonlinearities of the aerodynamic maps, remains unaffected. For the control design by TS-methods, the so-called c_Q - λ function shown in Fig. 3, where λ denotes the so-called tip-speed ratio

$$\lambda = \frac{\omega_r R}{v} = \frac{x_1 R}{v} \quad (5)$$

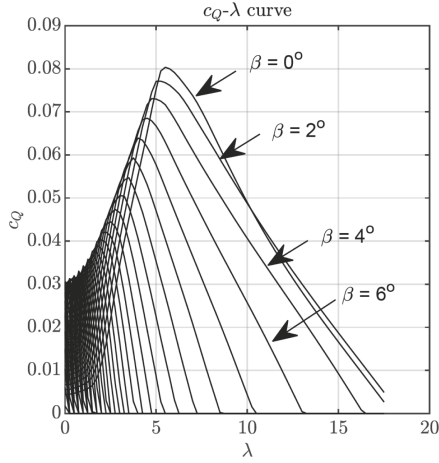


Fig. 3. c_Q - λ curve of NREL 5-MW wind turbine related to the collective rotor blade pitch angles $\beta \in \{0^\circ, 2^\circ, 4^\circ, \dots, 88^\circ, 90^\circ\}$

can be decomposed into

$$c_Q(\mathbf{x}, v) = c_{Q,1}(\mathbf{x}, v) + c_{Q,2}(\mathbf{x}, v). \tag{6}$$

The $c_{Q,i}$ functions are expressed as

$$c_{Q,i}(\lambda, \beta) = \tilde{c}_{Q,i}(\lambda, \beta) \frac{1 + \operatorname{sgn} \tilde{c}_Q(\lambda, \beta)}{2}, \quad i = 1, 2 \tag{7}$$

with

$$\begin{aligned} \tilde{c}_{Q,1}(\lambda, \beta) &= c_1 \left(1 + c_2 (\beta + c_3)^{\frac{1}{2}} \right) \\ &\quad + \frac{c_4}{\lambda} (c_5 \lambda_i(\lambda, \beta) - c_7 \beta^{c_8} - c_9) e^{(-c_{10} \lambda_i(\lambda, \beta))}, \\ \tilde{c}_{Q,2}(\lambda, \beta) &= -\frac{c_4}{\lambda} c_6 e^{(-c_{10} \lambda_i(\lambda, \beta))} \beta, \end{aligned} \tag{8}$$

where

$$\lambda_i(\lambda, \beta) = \frac{1}{\lambda + 0.08 \beta} - \frac{0.035}{c_{11} + c_{12} \beta^3}. \tag{9}$$

The coefficients c_j were determined by fitting (9) to a c_Q - λ curve calculated from the blade element theory obtained from FAST/AeroDyn simulations, see [16]. The following values were obtained:

$$\begin{array}{cccc} c_1 = 0.005 & c_2 = 1.53 & c_3 = 0.5 & c_4 = 0.18 \\ c_5 = 121 & c_6 = 27.9 & c_7 = 198 & c_8 = 2.36 \\ c_9 = 5.74 & c_{10} = 11.35 & c_{11} = 16.1 & c_{12} = 201. \end{array} \tag{10}$$

The variable λ in (7), (8), and (9) denotes the tip-speed ratio (5) and β is the pitch angle. Note, the necessity of the decomposition of $c_{Q,i}(\mathbf{x}, v)$ in (3) comes from the fact that the controllability of the $i = 1, \dots, N_r$ linear submodels ($\mathbf{A}_i, \mathbf{B}_i$) in the TS form

$$\dot{\mathbf{x}} = \sum_{i=1}^{N_r} h_i(\mathbf{z}) (\mathbf{A}_i \mathbf{x} + \mathbf{B}_i \mathbf{u}), \quad \mathbf{y} = \mathbf{C} \mathbf{x} \quad (11)$$

must be ensured for the controller synthesis. To convert a nonlinear model (3) into the TS model structure (11) the application-specific range of model validity is first specified by

$$\begin{aligned} \omega_r &\in [\underline{\omega}_r, \bar{\omega}_{r, \text{rated}}] = [0.8 \omega_{r, \text{rated}}, 1.4 \omega_{r, \text{rated}}], \\ \beta &\in [\underline{\beta}, \bar{\beta}] = [0, 0.45], \\ T_g &\in [\underline{T}_g, \bar{T}_g] = [0.2 T_{g, \text{rated}}, T_{g, \text{rated}}], \\ v &\in [10, 25], \end{aligned} \quad (12)$$

which is related to the permissible operating range with intervals of the rotor speed, blade pitch angle, and wind speed. Using that model validity range, an equivalent TS formulation of (11) is now obtained.

$$\begin{aligned} \dot{\mathbf{x}} &= \sum_{i=1}^{N_r=4} h_i(\mathbf{z}) \mathbf{A}_i \mathbf{x} + \mathbf{B} \mathbf{u}, \quad \mathbf{y} = \mathbf{C} \mathbf{x}, \\ \mathbf{z} &= (\omega_r, \beta, v, T_g)^T \end{aligned} \quad (13)$$

with

$$\begin{aligned} \mathbf{A}_1 &= \begin{pmatrix} f_1 & f_2 \\ 0 & -\frac{1}{\tau_\beta} \end{pmatrix}, \quad \mathbf{A}_2 = \begin{pmatrix} f_1 & \bar{f}_2 \\ 0 & -\frac{1}{\tau_\beta} \end{pmatrix}, \quad \mathbf{A}_3 = \begin{pmatrix} \bar{f}_1 & f_2 \\ 0 & -\frac{1}{\tau_\beta} \end{pmatrix}, \quad \mathbf{A}_4 = \begin{pmatrix} \bar{f}_1 & \bar{f}_2 \\ 0 & -\frac{1}{\tau_\beta} \end{pmatrix}, \\ \mathbf{B} &= \begin{pmatrix} 0 \\ \frac{1}{\tau_\beta} \end{pmatrix}, \quad \mathbf{C} = \begin{pmatrix} n_g & 0 \\ 0 & 1 \end{pmatrix} \end{aligned} \quad (14)$$

and the membership functions h_i of each i -th linear submodel $i = 1, \dots, N_r$. The relation defines the membership functions

$$\sum_{k=1}^{N_r} h_k(\mathbf{z}) = \prod_{j=1}^{N_l} \sum_{k=1}^2 w_{k,j}(\mathbf{z}), \quad N_r = 2^{N_l}, \quad (15)$$

where N_l denotes the number of nonlinear functions. Due to two nonlinear functions $N_l = 2$ in (3) we obtain $N_r = 2^{N_l} = 4$ membership functions

$$\begin{aligned} h_1(\mathbf{z}) &= w_{1,1}(\mathbf{z}) w_{1,2}(\mathbf{z}), & h_2(\mathbf{z}) &= w_{1,1}(\mathbf{z}) w_{2,2}(\mathbf{z}), \\ h_3(\mathbf{z}) &= w_{2,1}(\mathbf{z}) w_{1,2}(\mathbf{z}), & h_4(\mathbf{z}) &= w_{2,1}(\mathbf{z}) w_{2,2}(\mathbf{z}). \end{aligned} \quad (16)$$

The weighting functions result from the application of the sector linearity approach

$$\begin{aligned}
 w_{1,1} &= \frac{\bar{f}_1 - f_1(\mathbf{x}, T_g, v)}{\bar{f}_1 - \underline{f}_1}, & w_{2,1} &= \frac{f_1(\mathbf{x}, T_g, v) - \underline{f}_1}{\bar{f}_1 - \underline{f}_1}, \\
 w_{1,2} &= \frac{\bar{f}_2 - f_1(\mathbf{x}, v)}{\bar{f}_2 - \underline{f}_2}, & w_{2,2} &= \frac{f_2(\mathbf{x}, v) - \underline{f}_2}{\bar{f}_2 - \underline{f}_2}.
 \end{aligned}
 \tag{17}$$

with the min/max values of f_j

$$\bar{f}_j = \max_{\substack{\mathbf{x} \in [\underline{\mathbf{x}}, \bar{\mathbf{x}}] \\ T_g \in [T_g, \bar{T}_g] \\ v \in [v, \bar{v}]}} f_j, \quad \underline{f}_j = \min_{\substack{\mathbf{x} \in [\underline{\mathbf{x}}, \bar{\mathbf{x}}] \\ T_g \in [T_g, \bar{T}_g] \\ v \in [v, \bar{v}]}} f_j \quad \text{for } j = 1, 2.
 \tag{18}$$

The following numerical values are obtained for the 5-MW FAST reference turbine [12] using model parameters listed in the appendix of [16] :

$$\underline{f}_1 = -0.0942, \quad \bar{f}_1 = 0.3297, \quad \underline{f}_2 = -0.6290, \quad \bar{f}_2 = -0.1493.
 \tag{19}$$

Hence, the control-oriented model in TS form for the region above the design wind speed is clearly defined by (4), (13)–(19).

4 Observer Synthesis for Controller Reconfiguration

The observer design is based on a slightly different model than the model for the controller design (3). The model for the observer synthesis is valid in the entire range, i.e. for wind speeds below and above the design wind speed, and is given as follows

$$\begin{aligned}
 \dot{\bar{\mathbf{x}}} &= \begin{pmatrix} 0 & g_1(\bar{\mathbf{x}}) & g_2(\bar{\mathbf{x}}) \\ 0 & -\frac{1}{\tau_\beta} & 0 \\ 0 & 0 & -\frac{1}{\tau_v} \end{pmatrix} \bar{\mathbf{x}} + \begin{pmatrix} -\frac{n_g}{J} & 0 & 0 \\ 0 & \frac{1}{\tau_\beta} & 0 \\ 0 & 0 & \frac{1}{\tau_v} \end{pmatrix} \mathbf{u}, \\
 \mathbf{y} &= \begin{pmatrix} n_g & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \bar{\mathbf{x}},
 \end{aligned}
 \tag{20}$$

where

$$\begin{aligned}
 g_1(\bar{\mathbf{x}}) &= \frac{1}{2} \frac{\rho \pi R^3}{J} v^2 c_{Q,2}(\lambda(x_1, v), x_2) \frac{1}{x_2}, \\
 g_2(\bar{\mathbf{x}}) &= \frac{1}{2} \frac{\rho \pi R^3}{J} v c_{Q,1}(\lambda(x_1, v), x_2)
 \end{aligned}
 \tag{21}$$

with the state vector $\bar{\mathbf{x}} = (x_1, x_2, x_3)^T = (\omega_r, \beta, v)^T$ and the input vector $\mathbf{u} = (T_g, \beta_{ref}, v_{avg})^T$, where v_{avg} denotes the average wind speed, approximated from up

to 10 min of averaged time data measured on the downwind side of the rotor. For the fault reconstruction, two fault terms are added to (20) and formulated in TS form

$$\begin{aligned}\dot{\tilde{\mathbf{x}}} &= \sum_{i=1}^{N_r} h_i(\mathbf{z}) \bar{\mathbf{A}}_i \bar{\mathbf{x}} + \bar{\mathbf{B}} \bar{\mathbf{u}} + \bar{\mathbf{B}}_f \mathbf{f}, \\ \mathbf{y} &= \bar{\mathbf{C}} \bar{\mathbf{x}} + \bar{\mathbf{D}}_f \mathbf{f},\end{aligned}\quad (22)$$

where $N_r = 4$ with the fault vector $\mathbf{f} = (\mathbf{f}_a^T, \mathbf{f}_s^T)^T \in \mathbb{R}^n$ with $k = k_a + k_s$ containing the actuator faults $\mathbf{f}_a \in \mathbb{R}^{k_a}$ and sensor faults $\mathbf{f}_s \in \mathbb{R}^{k_s}$. The design idea for the proportional-multi-integral (PMI) observer proposed in [18] based on the introduction of an augmented state vector

$$\tilde{\mathbf{x}} = (\bar{\mathbf{x}}^T, \xi_1^T, \xi_2^T, \dots, \xi_q^T)^T, \quad (23)$$

where ξ is defined by the fault vector and its time derivatives

$$\xi_1 = \mathbf{f}^{(q-1)}, \quad \xi_2 = \mathbf{f}^{(q-2)}, \quad \dots, \quad \xi_{q-1} = \dot{\mathbf{f}}, \quad \xi_q = \mathbf{f}.$$

The q -th derivation of \mathbf{f} is assumed to be bounded. The objective of the PMI observer design is a robust state estimation with a dynamic fault reconstruction. The dynamics of the faults that can be occurred are considered by the q -th derivative of the fault signals divided into actuator and sensor faults. Here, $\bar{\mathbf{B}}_f$ and $\bar{\mathbf{D}}_f$ are chosen in such a way that the sensor faults affect the measurement equation, and the actuator faults affect the state differential equation in (22). After introducing (23), the design model for the PMI observer can be given as follows:

$$\begin{aligned}\dot{\tilde{\mathbf{x}}} &= \sum_{i=1}^{N_r} h_i(\mathbf{z}) \tilde{\mathbf{A}}_i \tilde{\mathbf{x}} + \tilde{\mathbf{B}} \mathbf{u} + \tilde{\mathbf{G}} \mathbf{f}^{(q)}, \\ \mathbf{y} &= \tilde{\mathbf{C}} \tilde{\mathbf{x}},\end{aligned}\quad (24)$$

where

$$\begin{aligned}\tilde{\mathbf{A}}_i &= \begin{pmatrix} \bar{\mathbf{A}}_i & \mathbf{0}_{n \times k(q-1)} & \mathbf{B}_{f,n \times k} \\ & \mathbf{0}_{k \times (n+kq)} & \\ \mathbf{0}_{k(q-1) \times n} & \mathbf{I}_{k(q-1)} & \mathbf{0}_{k(q-1) \times k} \end{pmatrix}, \\ \tilde{\mathbf{C}} &= (\bar{\mathbf{C}}_{p \times n}, \mathbf{0}_{p \times k(q-1)}, \mathbf{D}_{f,p \times k}), \\ \tilde{\mathbf{B}} &= \begin{pmatrix} \bar{\mathbf{B}} \\ \mathbf{0}_{kq \times m} \end{pmatrix}, \quad \tilde{\mathbf{G}} = \begin{pmatrix} \mathbf{0}_{n \times k} \\ \mathbf{I}_k \\ \mathbf{0}_{k(q-1) \times k} \end{pmatrix}.\end{aligned}$$

The PMI observer law derived from the previous model (24) in the form of a Luenberger observer is given by

$$\dot{\hat{\tilde{\mathbf{x}}}} = \sum_{i=1}^{N_r} h_i(\mathbf{z}) \left(\tilde{\mathbf{A}}_i \hat{\tilde{\mathbf{x}}} + \tilde{\mathbf{L}}_i (\mathbf{y} - \hat{\mathbf{y}}) \right) + \tilde{\mathbf{B}} \mathbf{u} \quad (25)$$

with $y_3 = v_{obs}$ as the estimated wind speed of a series-connected wind speed observer (cascaded observer concept) with the related state error equation

$$\dot{\tilde{\mathbf{e}}} = \sum_{i=1}^{N_r} h_i(\mathbf{z}) (\tilde{\mathbf{A}}_i - \tilde{\mathbf{L}}_i \tilde{\mathbf{C}}) \tilde{\mathbf{e}} - \tilde{\mathbf{G}} \mathbf{f}^{(q)}, \quad \tilde{\mathbf{e}} = \hat{\tilde{\mathbf{x}}} - \tilde{\mathbf{x}}. \quad (26)$$

To estimate both actuator faults, the generator torque, and the pitch angle (see Fig. 1) by the PMI observer (25), the faults dynamics to be reconstructed must be specified. Good results related to a short reconstruction time with little overshoot could be achieved with a “depth of derivation” of $q = 2$ for $k = 2$ which results from $k = k_a + k_s$ with $k_a = 2$ and $k_s = 0$. For the presented design, the augmented vector (23) is then

$$\dot{\tilde{\mathbf{x}}} = (\bar{\mathbf{x}}^T, \dot{\mathbf{f}}^T, \mathbf{f}^T)^T \quad (27)$$

with $\mathbf{f} = (f_{a,Tg}, f_{a,\beta})^T$, where $f_{a,Tg}$ denotes the generator fault and $f_{a,\beta}$ the pitch actuator fault. The corresponding fault distribution matrices are given by

$$\mathbf{B}_f = \begin{pmatrix} -\frac{n_g}{J} & 0 \\ 0 & \frac{1}{\tau_\beta} \\ 0 & 0 \end{pmatrix}, \quad \mathbf{D}_f = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}. \quad (28)$$

The actuator fault estimate by (25) is added to the output \mathbf{u} of the nominal controller .

$$\mathbf{u}_{FTC} = \mathbf{u} + \hat{\mathbf{f}}_a. \quad (29)$$

Note that this is the only calculation done in the reconfiguration block of Fig. 1. The observer design is based on the procedure according to [19]. First, a Lyapunov-based design inspired by [18] is used for each local submodel. If the pair $(\tilde{\mathbf{A}}_i, \tilde{\mathbf{C}})$ is completely observable and let $\mathbf{X} > 0$ be a solution of the Lyapunov equation

$$(\tilde{\mathbf{A}}_i + \mu \mathbf{I})^T \mathbf{X}_i + \mathbf{X}_i (\tilde{\mathbf{A}}_i + \mu \mathbf{I}) = 2 \tilde{\mathbf{C}}^T \tilde{\mathbf{C}} \quad (30)$$

with

$$-\mu < \Re\{\lambda_{min}(\tilde{\mathbf{A}}_i)\} \quad (\mu > 0), \quad (31)$$

then the matrix $\tilde{\mathbf{A}}_i - \tilde{\mathbf{L}}_i \tilde{\mathbf{C}}$ is stable if the observer gain matrix is calculated by

$$\tilde{\mathbf{L}}_i = \mathbf{X}_i^{-1} \tilde{\mathbf{C}}^T. \quad (32)$$

The error system (26) is stable with the exponential decay rate 2μ of the zero-input response. In the second step, the stability of the error equation of the weighted combinations of linear submodels has to be ensured by the LMI-based stability conditions proposed in [19]. The error system (26) with the assumption that $\mathbf{f}^{(q)} = \mathbf{0}$ and $\tilde{\mathbf{e}}(t=0) \neq \mathbf{0}$ is asymptotic stable if there exists a common $\mathbf{P} = \mathbf{P}^T$, $\mathbf{P} > 0$, and $\tilde{\mathbf{L}}_i$, $i = 1, 2, \dots, N_r$ if the LMIs

$$\mathbf{P} (\tilde{\mathbf{A}}_i - \tilde{\mathbf{L}}_i \tilde{\mathbf{C}}) + (\tilde{\mathbf{A}}_i - \tilde{\mathbf{L}}_i \tilde{\mathbf{C}})^T \mathbf{P} < 0 \quad (33)$$

hold for $i = 1, \dots, N_r$. The proof is given in [19].

5 Power Tracking Controller Validation with FTC

To validate the FTC for power tracking, a fault was applied to the pitch system. For this purpose, a sawtooth-shaped fault was investigated at $t = 32$ s and stopped at $t = 72$ s. The fault and its reconstruction are shown in Fig. 4. One can see the reconstruction quality of this. However, a large but also rapidly decaying estimation error can be seen in the reversal points of the fault curve at $t = 40$ s, 48 s, 56 s, 63 s, and 72 s.

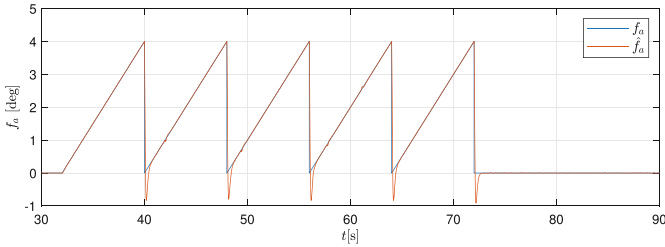


Fig. 4. Fault reconstruction of a pitch actuator fault with a sawtooth shape during $32 \text{ s} \leq t \leq 72 \text{ s}$

How the entire control system without FTC reacts to that actuator fault in the pitch system with active power control can be seen in Fig. 5 in the left diagram. The intermittent pitch actuator fault with a significant amplitude range causes a strong fluctuation in the rotor speed. This results in strong power fluctuations, which the nominal controller

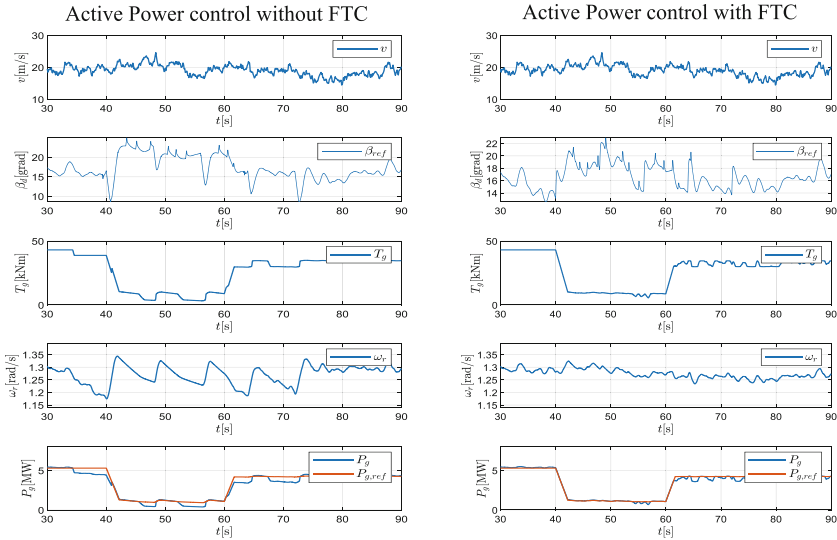


Fig. 5. Comparison of power tracking behavior without and with FTC during pitch actuator fault occurring $32 \text{ s} \leq t \leq 72 \text{ s}$ shown in Fig. 4

cannot smooth. The generated power output is shown in the bottom diagram on the left. Comparing this curve with the generated generator power on the right side of Fig. 5, one can easily distinguish the significant improvement using FTC. The turbine can largely follow the reference value specification $P_{g,ref}$ of the higher-level system. The absolute powers are shown for better comparability and not the required differential power. In addition to the good performance tracking utilizing generator torque adjustment, the FTC's good setpoint control of the rotor speed can also be noted.

6 Conclusion

This paper proposes a fault-tolerant control system for wind turbines with power tracking capability above the rated wind speed. The control system consists of a nominal controller and a reconfiguration block. A proportional-multi-integral observer reconstructs the actuator fault, which will be superimposed on the nominal actuator signal in the reconfiguration block. An analytical model-based design method for the reconfiguration block is presented in detail. System simulations are used to show how this will improve the reliability of future power-tracking wind turbines. In future work, the control concept will be integrated into a grid control system to investigate the influence on the stability and performance of the higher-level power system. The aim is to improve the resilience of power systems with a high proportion of wind farms with power tracking capability under the fault influence of the sub-components at the power plant level.

Acknowledgment. This research is part of the EU-Project POSYTYF (POwering SYstem flexiBiliTY in the Future through RES), <https://posytyf-h2020.eu>. The POSYTYF project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 883985.

References

1. Kusche, S., Pöschke, F., Schulte, H.: Objectives and assessment criteria for controller design. In: POSYTYF (H2020 Project) - Deliverable 2.2, Tech. Report (2022). <https://posytyf-h2020.eu/english-version/deliverables>
2. van Kuik, G.A.M., et al.: Long-term research challenges in wind energy - a research agenda by the European Academy of wind energy. *Wind Energy Science* **1**(1), 1–39 (2016)
3. Das, K., Altin, M., Hansen, A., Sørensen, P., Flynn, D., Abildgaard, H.: Wind power support during overfrequency emergency events. *CIGRE Sci. Eng.* **9**, 73–83 (2018)
4. Aho, J., Fleming, P., Pao, L.Y.: Active power control of wind turbines for ancillary services: a comparison of pitch and torque control methodologies. In: American Control Conference (ACC), Boston, MA, USA (2016)
5. Galinos, C., Urban, A.M., Lio, W.H.: Optimised de-rated wind turbine response and loading through extended controller gain-scheduling, *J. Phys. Conf. Ser.* **1222**(012020), 1–8 (2019)
6. Shan, M., Shan, W., Welck, F., Duckwitz, D.: Design and laboratory test of black-start control mode for wind turbines. *Wind Energy* **23**(3), 763–778 (2020)
7. Ibáñez, B., Inthamoussou, F.: De Battista H (2020) Wind turbine load analysis of a full range LPV controller. *Renewable Energy* **145**, 2741–2753 (2020)

8. Inthamoussou, F.A., Battista, H.D., Mantz, R.J.: LPV-based active power control of wind turbines covering the complete wind speed range. *Renewable Energy* **99**, 996–1007 (2016)
9. Pöschke, F., Fortmann J., Schulte, H.: Nonlinear wind turbine controller for variable power generation in full load region. In: American Control Conference, Sheraton Hotel, Seattle, USA, 2017, pp. 1395–1400 (2017)
10. Pöschke, F., Gauterin, E., Kühn, M., Fortmann, J., Schulte, H.: Load mitigation and power tracking capability for wind turbines using linear matrix inequality-based control design. *Wind Energy* **23**(9), 1792–1809 (2020)
11. Kim, K., Kim, H.G., Kim, C.J., Paek, I., Bottasso, C.L., Campagnolo, F.: Design and validation of demanded power point tracking control algorithm of wind turbine. *Int. J. Prec. Eng. Manufact.–Green Technol.* **5**, 387–400 (2018)
12. Jonkman, J.M., Butterfield, S., Musial, W., Scott, G.: Definition of a 5-MW Reference Wind Turbine for Offshore System Development. Tech, Report (2009)
13. Pöschke, F., Petrović, V., Berger, F., Hölling, L.N.M., Kühn, M., Schulte, H.: Model-based wind turbine control design with power tracking capability: a wind-tunnel validation. *Control. Eng. Pract.* **120**(105014), 1–13 (2022)
14. Steffen, T.: Control Reconfiguration of Dynamical Systems. Springer (2005). <https://doi.org/10.1007/b107072>
15. Blanke, M., Kinnaert, M., Lunze, J., Staroswiecki, M.: Diagnosis and Fault-Tolerant Control (2nd ed.), Springer (2006). <https://doi.org/10.1007/978-3-540-35653-0>
16. Georg, S.: Fault diagnosis and fault-tolerant control of wind turbines - Nonlinear Takagi-Sugeno and sliding mode techniques, Ph. D. Thesis, HTW Berlin, Control Engineering Group / University Rostock, Fakultät für Maschinenbau und Schiffstechnik (2015)
17. Li, S., Wang, H., Aitouche, A., Tian, Y., Christov, N.: Active fault tolerance control of a wind turbine system using an unknown input observer with an actuator fault. *Int. J. Appl. Math. Comput. Sci.* **32**(1), 69–81 (2018)
18. Gao, Z., Ding, S.X., Ma, Y.: Robust fault estimation approach and its application in vehicle lateral dynamic systems. *Optim. Control Appl. Methods* **28**, 143–156 (2007). <https://doi.org/10.1002/oca.786>
19. Kühne, P., Pöschke, F., Schulte, H.: Fault estimation and fault-tolerant control of the FAST NREL 5-MW reference wind turbine using a proportional multi-integral observer. *Int. J. Adapt. Control Signal Process.* **32**(4), 568–585 (2018). <https://doi.org/10.1002/acs.2800>



A Multiple Actuator and Sensor Fault Estimation for Dynamic Systems

Marcin Pazera^(✉) , Marcin Witczak , and Józef Korbicz 

Institute of Control and Computation Engineering, University of Zielona Góra,
ul. Szafrana 2, 65-516 Zielona Góra, Poland
{m.pazera,m.witczak,j.korbicz}@issi.uz.zgora.pl

Abstract. The paper deals with the problem of simultaneous actuator and sensor fault estimation for linear dynamic systems on which influence unknown exogenous disturbances. The most common approach of dealing with fault estimation approach is that either actuator or sensor fault estimation schemes are provided in the literature. This paper proposes a kind of a fusion of those approaches which results with the capability of estimation of actuator and sensor faults which might appear in the system simultaneously. The novelty relies on that the unwelcome rate of change of the fault factor is vanished which simplifies the design problem. The methodology utilized in the paper to attain the robustness is the so-called Quadratic boundedness approach. In the final part of the paper, the illustrative example with the implementation to the laboratory Multi-Tank system is provided which shows the performance of the proposed approach.

Keywords: Fault diagnosis · fault estimation · multiple faults · actuator and sensor fault · robust estimation

1 Introduction

A key concept behind a desired performance of dynamic processes pertains to closed-loop feedback control. That a strategy allows applying a controller to a plant in such a way that the controlled system would be guided in a desired manner by suitably reacting to changes of an output of the system. Starting from widespread PIDs, in a continuous or in a discrete-time domain [3, 18], through the classical state- or output feedback [7, 17] approach, there can be found more sophisticated solutions as, e.g. ILC (Iterative Learning Control) [9, 19], which principally is used for repetitive processes where a device continuously performs the same actions. However, a core of a reliable control is devoted to a so-called FTC (Fault-Tolerant Control) [8, 28]. FTC is applied to plants where a high safety conditions are required and the system should be guided safely and as close as possible to a reference even if some faults occur. Additionally, robust solutions have been also developed for FTC [2, 25] which results in a safely operating system under the presence of faults with simultaneous robustness against disturbances and noises. A crucial aspect in a FTC is to provide a suitable control law to a system which might have faults and thereby enhance the utility of

that system. Obviously, the faults acting onto the system may, and in many cases do give rise to performance degradation thus, such a FTC approach is responsible to bring the system back to its nominal operating point. There are two concepts to attain the fault tolerance, namely a passive and active FTC [10, 14]. Such a division entails two ways of achieving fault tolerance in opposite manner, which use different, fixed or reconfigurable control strategies [15, 29].

However, before designing the FTC scheme, an information about fault that might act the system is obviously required. A fault estimation theory constitutes an answer to cope such requirements. Fault estimation approaches commonly proposed in literature concern either actuator or sensor (c.f [12, 16, 20, 21]) fault estimation. In other words the exact one kind of fault is possible to be estimated. However, taking into account the fact that surrounding industrial environment is rather more intricate, it cannot be allowed for such a lack. Especially nowadays we cannot imagine any system without either sensors or actuators. They always operate together and cooperate with each other in some extent. The question that arises here is what will happen if the sensor fault occurs while estimating the actuator fault only? On the other hand, what will happen if the actuator fault occurs in case of estimating the sensor fault only? In such cases the approaches for estimating either actuator or sensor fault neglecting a part of an entire system, which are widely investigated in the literature, are not sufficient and for this reason the results may be incorrect. Instead, the approach proposed in this paper is a solution towards estimating both actuator and sensor faults. Indeed, there will be no more assumptions concerning which kind of fault should be estimated. For such a reason there is a necessity to develop an approach which can handle that challenge. In this paper, Sect. 2 formulates the problem while in the Sect. 3 a fault estimator capable for recovering both actuator and sensor fault is proposed. In Sect. 4 an illustrative example with implementation to Multi-Tank system is shown and finally, Sect. 5 concludes the paper.

2 Preliminaries

In order to deal with simultaneous estimation of actuator and sensor faults, let us take into consideration the following system:

$$\mathbf{x}_{k+1} = \mathbf{A}\mathbf{x}_k + \mathbf{B}\mathbf{u}_k + \mathbf{L}_u\mathbf{f}_{u,k} + \mathbf{D}_1\mathbf{d}_{x,k}, \quad (1)$$

$$\mathbf{y}_k = \mathbf{C}\mathbf{x}_k + \mathbf{L}_y\mathbf{f}_{y,k} + \mathbf{D}_2\mathbf{d}_{y,k}, \quad (2)$$

where $\mathbf{x}_k = [x_1, x_2, \dots, x_n] \in \mathbb{R}^n$, $\mathbf{u}_k = [u_1, u_2, \dots, u_r] \in \mathbb{R}^r$, $\mathbf{y}_k = [y_1, y_2, \dots, y_m] \in \mathbb{R}^m$, represent the state, control input and measured output vectors, respectively. Additionally, $\mathbf{f}_{u,k} = [f_{u,1}, f_{u,2}, \dots, f_{u,n_u}] \in \mathbb{R}^{n_u}$ states for the fault vector which acts directly onto the state vector \mathbf{x}_k and throughout this work it will be referred to an *actuator fault*, whilst $\mathbf{f}_{y,k} = [f_{y,1}, f_{y,2}, \dots, f_{y,n_y}] \in \mathbb{R}^{n_y}$ is the fault vector which affects the measured output and for such a reason it will be referred to a *sensor fault*. Moreover, $\mathbf{d}_{x,k} = [d_{x,1}, d_{x,2}, \dots, d_{x,q_1}] \in \mathbb{R}^{q_1}$ and $\mathbf{d}_{y,k} = [d_{y,1}, d_{y,2}, \dots, d_{y,q_2}] \in \mathbb{R}^{q_2}$ are unknown exogenous process uncertainty and exogenous measurement noise vectors, respectively, where k stands

for a discrete-time instance. Moreover, matrices \mathbf{L}_u and \mathbf{L}_y denote actuator and sensor fault distribution matrices, respectively. In the other words, they describe the way how the appropriate fault vector influences the system.

The only one assumption or preferably limitation is the number of faults that can be estimated which is rather obvious, because it is not possible to estimate more faults than a number of measurements are read from the system. Thus, this limitation is contained in the relation that $n_u + n_y = m$.

3 An Estimation Scheme for Multiple Faults

A following approach is proposed to be able to estimate the actuator and sensor faults and a state itself, simultaneously. To sort out such an issue an estimator of the following structure is to be investigated:

$$\hat{\mathbf{f}}_{y,k} = \mathbf{E}\mathbf{y}_k - \bar{\mathbf{A}}\hat{\mathbf{x}}_{k-1} - \bar{\mathbf{B}}\mathbf{u}_{k-1} - \bar{\mathbf{L}}_u\hat{\mathbf{f}}_{u,k-1}, \quad (3)$$

$$\hat{\mathbf{x}}_{k+1} = \mathbf{A}\hat{\mathbf{x}}_k + \mathbf{B}\mathbf{u}_k + \mathbf{L}_u\hat{\mathbf{f}}_{u,k} + \mathbf{F}_x(\mathbf{y}_k - \hat{\mathbf{y}}_k), \quad (4)$$

$$\hat{\mathbf{f}}_{u,k+1} = \hat{\mathbf{f}}_{u,k} + \mathbf{F}_u(\mathbf{y}_k - \hat{\mathbf{y}}_k), \quad (5)$$

$$\hat{\mathbf{y}}_k = \mathbf{C}\hat{\mathbf{x}}_k + \mathbf{L}_y\hat{\mathbf{f}}_{y,k}, \quad (6)$$

where \mathbf{F}_x as well as \mathbf{F}_u are the state and actuator fault estimation gain matrices of appropriate dimensions, respectively, whilst $\bar{\mathbf{A}} = \bar{\mathbf{E}}\mathbf{A}$, $\bar{\mathbf{B}} = \bar{\mathbf{E}}\mathbf{B}$, $\bar{\mathbf{L}}_u = \bar{\mathbf{E}}\mathbf{L}_u$, $\bar{\mathbf{D}}_1 = \bar{\mathbf{E}}\mathbf{D}_1$ and $\bar{\mathbf{E}} = \mathbf{E}\mathbf{C}$. The matrix \mathbf{E} means a pseudo-inversion of the sensor fault distribution matrix \mathbf{L}_y such that $\mathbf{E} = (\mathbf{L}_y)^\dagger$, $\mathbf{E}\mathbf{L}_y = \mathbf{I}$. Such an approach constitutes a kind of a fusion of the approaches proposed and presented in the literature. Of course, there is no exact combination of those approaches, due to observability aspects, though they have been suitably adapted to cope the requirements. As a result there is a possibility to reconstruct the state as well as both actuator and sensor faults.

As the means of obtaining these gain matrices, the robust QB approach is applied. To do so, let us start by receiving an estimation error of the state. Having regarded (1)–(2) as well as (3)–(6) the dynamics of the state estimation error obeys

$$\mathbf{e}_{x,k+1} = [\mathbf{A} - \mathbf{F}_x\mathbf{C}] \mathbf{e}_{x,k} + \mathbf{L}_u\mathbf{e}_{u,k} - \mathbf{F}_x\mathbf{L}_y\mathbf{e}_{y,k} + \mathbf{D}_1\mathbf{d}_{x,k} - \mathbf{F}_x\mathbf{D}_2\mathbf{d}_{y,k}, \quad (7)$$

where $\mathbf{e}_{u,k}$ and $\mathbf{e}_{y,k}$ denote an actuator fault estimation error and a sensor fault estimation error, respectively. The next one is the dynamics of the actuator fault estimation error which is given by

$$\mathbf{e}_{u,k+1} = \boldsymbol{\varepsilon}_{u,k} + \mathbf{e}_{u,k} - \mathbf{F}_u\mathbf{C}\mathbf{e}_{x,k} - \mathbf{F}_u\mathbf{L}_y\mathbf{e}_{y,k} - \mathbf{F}_u\mathbf{D}_2\mathbf{d}_{y,k}, \quad (8)$$

with $\boldsymbol{\varepsilon}_{u,k} = \mathbf{f}_{u,k+1} - \mathbf{f}_{u,k}$ denoting the error between sequentially following samples of the actuator fault. Now, having such defined two of the estimation errors it is time to establish the dynamics of the sensor fault estimation error. Although, it is necessary to transform the output equation (2) in a way to be

able to receive the sensor fault equation of the form

$$\mathbf{f}_{y,k} = \mathbf{E}\mathbf{y}_k - \bar{\mathbf{A}}\mathbf{x}_{k-1} - \bar{\mathbf{B}}\mathbf{u}_{k-1} - \bar{\mathbf{L}}_u\mathbf{f}_{u,k-1} - \bar{\mathbf{D}}_1\mathbf{d}_{x,k-1} - \mathbf{E}\mathbf{D}_2\mathbf{d}_{y,k}. \quad (9)$$

Hence, taking into consideration the sensor fault (9) as well as its estimate (3), the estimation error of this particular fault can be presented as follows

$$\mathbf{e}_{y,k} = \mathbf{f}_{y,k} - \hat{\mathbf{f}}_{y,k} = -\bar{\mathbf{A}}\mathbf{e}_{x,k-1} - \bar{\mathbf{L}}_u\mathbf{e}_{u,k-1} - \bar{\mathbf{D}}_1\mathbf{d}_{x,k-1} - \mathbf{E}\mathbf{D}_2\mathbf{d}_{y,k}. \quad (10)$$

It can be readily observed that the sensor fault estimation error (10) does not depend on itself, in the other words the previous samples of this error do not influence $\mathbf{e}_{y,k}$. On the other hand, it depends on the previous samples of both, state $\mathbf{e}_{x,k-1}$ as well as actuator fault $\mathbf{e}_{u,k-1}$ estimation errors. Since, the dynamics of the sensor fault estimation error is established, for further analysis, let (10) be applied to the state (7) and actuator fault (8) estimation errors, which yield

$$\begin{aligned} \mathbf{e}_{x,k+1} = & (\mathbf{A} - \mathbf{F}_x\mathbf{C})\mathbf{e}_{x,k} + \mathbf{F}_x\mathbf{L}_y\bar{\mathbf{A}}\mathbf{e}_{x,k-1} + \mathbf{L}_u\mathbf{e}_{u,k} + \mathbf{F}_x\mathbf{L}_y\bar{\mathbf{L}}_u\mathbf{e}_{u,k-1} \\ & + \mathbf{F}_x\mathbf{L}_y\mathbf{D}_1\mathbf{d}_{x,k-1} + \mathbf{D}_1\mathbf{d}_{x,k} + (\mathbf{F}_x\mathbf{L}_y\mathbf{E}\mathbf{D}_2 - \mathbf{F}_x\mathbf{D}_2)\mathbf{d}_{y,k}, \end{aligned} \quad (11)$$

$$\begin{aligned} \mathbf{e}_{u,k+1} = & \boldsymbol{\varepsilon}_{u,k} + \mathbf{e}_{u,k} - \mathbf{F}_u\mathbf{C}\mathbf{e}_{x,k} + \mathbf{F}_u\mathbf{L}_y\bar{\mathbf{A}}\mathbf{e}_{x,k-1} + \mathbf{F}_u\mathbf{L}_y\bar{\mathbf{L}}_u\mathbf{e}_{u,k-1} \\ & + \mathbf{F}_u\mathbf{L}_y\bar{\mathbf{D}}_1\mathbf{d}_{x,k-1} + (\mathbf{F}_u\mathbf{L}_y\mathbf{E}\mathbf{D}_2 - \mathbf{F}_u\mathbf{D}_2)\mathbf{d}_{y,k}. \end{aligned} \quad (12)$$

The structure of the equation capable of estimating the sensor fault has allowed us to reduce the sensor fault estimation error without unwelcome factor concerning the rate of change of sensor fault. Unfortunately, such a factor is removed only in case of sensor fault, though it has not been possible to achieve such an outcome in case of actuator fault. Of course, there are in the literature such approaches in which the rate of change of actuator fault factor is removed in case of reconstructing this particular fault [21,27] or even in case of estimating both [22,23], however, such approaches cannot be applied directly to the desirable fault-tolerant control scheme due to the structure of the estimator in those cases. In the other words, in the literature indeed, both factors from actuator and sensor faults are removed but the scheme used in those approaches is not applicable because of the use of output equation assumed to be in future instances of time, i.e., $k+1$. Even though, that approach eliminates the rate of change of the actuator fault, however, it provides more complicated structure of the estimator and requires the system being converted into descriptor-like [13]. Moreover, in the proposed in this work procedure of the estimation, the sensor fault estimation error is vanished from the state estimation error (11) as well as actuator fault estimation error (12). The one proposed in this paper indeed does not employ signals at time $k+1$, although the variable are delayed of a sample.

The definitions of such dynamics of estimation errors allow us for introducing new temporary variables build of the estimation errors of both state and actuator fault, and also build with disturbances and the rate of change of actuator fault factor:

$$\tilde{\mathbf{e}}_k = [\mathbf{e}_{x,k}^T \ \mathbf{e}_{u,k}^T]^T, \quad (13)$$

$$\tilde{\mathbf{d}}_k = [\mathbf{d}_{x,k}^T \ \mathbf{d}_{y,k}^T \ \boldsymbol{\varepsilon}_{u,k}^T]^T. \quad (14)$$

It can be pointed out that the sensor fault estimation error is not contained in (13). This is achieved because the sensor fault estimation is obtained directly from the output equation, and there is no need to design an additional gain matrix required to estimate this specific fault. Thus, bearing them in mind, it is feasible to put the state estimation error as well as actuator fault estimation error compactly into a following form

$$\tilde{e}_{k+1} = \tilde{X}_1 \tilde{e}_k + \tilde{X}_2 \tilde{e}_{k-1} + \tilde{Z}_1 \tilde{d}_k + \tilde{Z}_2 \tilde{d}_{k-1}, \tag{15}$$

where $\tilde{X}_1 = \tilde{A}_1 - \tilde{F}\tilde{C}$, $\tilde{X}_2 = \tilde{F}\tilde{A}_2$, $\tilde{Z}_1 = \tilde{D}_1 - \tilde{F}\tilde{D}_2$ and $\tilde{Z}_2 = \tilde{F}\tilde{D}_3$ with

$$\begin{aligned} \tilde{A}_1 &= \begin{bmatrix} A & L_u \\ 0 & I \end{bmatrix}, & \tilde{A}_2 &= [L_y \bar{A} \ L_y \bar{L}_u], & \tilde{C} &= [C \ 0], \\ \tilde{D}_1 &= \begin{bmatrix} D_1 & 0 & 0 \\ 0 & 0 & I \end{bmatrix}, & \tilde{D}_2 &= [0 \ L_y E D_2 - D_2 \ 0], & \tilde{F} &= \begin{bmatrix} F_x \\ F_u \end{bmatrix}, \\ & & \tilde{D}_3 &= [L_y D_1 \ 0 \ 0]. \end{aligned} \tag{16}$$

Remark 1. It is worth to remember that the employed Quadratic Boundedness [1, 5, 6] approach has been chosen to achieve a desired robust stability of the estimator. Initially, such an approach was applied in order to design a state estimator for linear, uncertain, discrete-time systems, although an extension provided within this work allow for applying it to the proposed ASFE guaranteeing the robust stability of this particular estimator able to estimate the state of the system as well as both, actuator and sensor faults, simultaneously.

Let the Lyapunov candidate function be described as

$$V_k = \tilde{e}_k^T \mathbf{R} \tilde{e}_k + \tilde{e}_{k-1}^T \mathbf{S} \tilde{e}_{k-1}, \tag{17}$$

with $\mathbf{R} \succ 0$ and $\mathbf{S} \succ 0$.

Based upon the variables previously defined, denoting estimation error \tilde{e}_k for the state and actuator fault, the following theorem sum up the estimator design:

Theorem 1. *The system given by (15) is strictly quadratically bounded for all $\tilde{d}_k \in \mathbb{E}_d$ if there exist matrices $\mathbf{R} \succ 0$, $\mathbf{S} \succ 0$ and \mathbf{N} as well as scalars $\alpha \in (0, 1)$ and $\beta \in (0, 1)$ like that $\alpha + \beta < 1$, such that the following condition is met*

$$\begin{bmatrix} \mathbf{S} - (1 - \alpha - \beta) \mathbf{R} & * & * & * & * \\ 0 & -(1 - \alpha - \beta) \mathbf{S} & * & * & * \\ 0 & 0 & -\alpha \mathbf{Q}_d & * & * \\ 0 & 0 & 0 & -\beta \mathbf{Q}_d & * \\ \mathbf{R} \tilde{A}_1 - \mathbf{N} \tilde{C} & \mathbf{N} \tilde{A}_2 & \mathbf{R} \tilde{D}_1 - \mathbf{N} \tilde{D}_2 & \mathbf{N} \tilde{D}_3 & -\mathbf{R} \end{bmatrix} \prec 0. \tag{18}$$

where $*$ stands for transpose of an appropriate element of the symmetric matrix.

Proof. As it has already been stated, the compact form of the estimator (15) stands for a single delay system, thus a following invariant set is proposed

$$\mathbb{E}_d = \left\{ (\tilde{e}_k, \tilde{e}_{k-1}) : \tilde{e}_k^T \mathbf{R} \tilde{e}_k + \tilde{e}_{k-1}^T \mathbf{S} \tilde{e}_{k-1} \leq 1 \right\}. \tag{19}$$

Now, let us recall the following definition:

Definition 1. *The system indicated by (1)–(2) is strictly quadratically bounded for all $\tilde{\mathbf{d}}_k \in \mathbb{E}_{d,k} \geq 0$ if $V_k > 1 \Rightarrow V_{k+1} - V_k < 0$ for any $\tilde{\mathbf{d}}_k \in \mathbb{E}_{d,k}$.*

Thus, employing the definition 1 along with the facts that $\tilde{\mathbf{d}}_k^T \mathbf{Q}_d \tilde{\mathbf{d}}_k \leq 1$ as well as $\tilde{\mathbf{d}}_{k-1}^T \mathbf{Q}_d \tilde{\mathbf{d}}_{k-1} \leq 1$, it can be shown that:

$$\tilde{\mathbf{d}}_{k-1}^T \mathbf{Q}_d \tilde{\mathbf{d}}_{k-1} < \tilde{\mathbf{e}}_k^T \mathbf{R} \tilde{\mathbf{e}}_k + \tilde{\mathbf{e}}_{k-1}^T \mathbf{S} \tilde{\mathbf{e}}_{k-1}, \tag{20}$$

$$\tilde{\mathbf{d}}_k^T \mathbf{Q}_d \tilde{\mathbf{d}}_k < \tilde{\mathbf{e}}_k^T \mathbf{R} \tilde{\mathbf{e}}_k + \tilde{\mathbf{e}}_{k-1}^T \mathbf{S} \tilde{\mathbf{e}}_{k-1}. \tag{21}$$

Then, by establishing a new temporary variable $\mathbf{w}_k = \left[\tilde{\mathbf{e}}_k^T \ \tilde{\mathbf{e}}_{k-1}^T \ \tilde{\mathbf{d}}_k^T \ \tilde{\mathbf{d}}_{k-1}^T \right]^T$, it can be easily shown that

$$\mathbf{w}_k^T \begin{bmatrix} \tilde{\mathbf{X}}_1^T \mathbf{R} \tilde{\mathbf{X}}_1 + \mathbf{S} - \mathbf{R} & \tilde{\mathbf{X}}_1^T \mathbf{R} \tilde{\mathbf{X}}_2 & \tilde{\mathbf{X}}_1^T \mathbf{R} \tilde{\mathbf{Z}}_1 & \tilde{\mathbf{X}}_1^T \mathbf{R} \tilde{\mathbf{Z}}_2 \\ \tilde{\mathbf{X}}_2^T \mathbf{R} \tilde{\mathbf{X}}_1 & \tilde{\mathbf{X}}_2^T \mathbf{R} \tilde{\mathbf{X}}_2 - \mathbf{S} & \tilde{\mathbf{X}}_2^T \mathbf{R} \tilde{\mathbf{Z}}_1 & \tilde{\mathbf{X}}_2^T \mathbf{R} \tilde{\mathbf{Z}}_2 \\ \tilde{\mathbf{Z}}_1^T \mathbf{R} \tilde{\mathbf{X}}_1 & \tilde{\mathbf{Z}}_1^T \mathbf{R} \tilde{\mathbf{X}}_2 & \tilde{\mathbf{Z}}_1^T \mathbf{R} \tilde{\mathbf{Z}}_1 & \tilde{\mathbf{Z}}_1^T \mathbf{R} \tilde{\mathbf{Z}}_2 \\ \tilde{\mathbf{Z}}_2^T \mathbf{R} \tilde{\mathbf{X}}_1 & \tilde{\mathbf{Z}}_2^T \mathbf{R} \tilde{\mathbf{X}}_2 & \tilde{\mathbf{Z}}_2^T \mathbf{R} \tilde{\mathbf{Z}}_1 & \tilde{\mathbf{Z}}_2^T \mathbf{R} \tilde{\mathbf{Z}}_2 \end{bmatrix} \mathbf{w}_k < 0. \tag{22}$$

Taking into account (20)–(21), it is obvious that for $\alpha > 0$ and $\beta > 0$:

$$\alpha \mathbf{w}_k^T \begin{bmatrix} -\mathbf{R} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & -\mathbf{S} & \mathbf{0} & \mathbf{0} \\ -\mathbf{R} & \mathbf{0} & \mathbf{Q}_d & \mathbf{0} \\ -\mathbf{R} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{w}_k < 0, \tag{23}$$

$$\beta \mathbf{w}_k^T \begin{bmatrix} -\mathbf{R} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & -\mathbf{S} & \mathbf{0} & \mathbf{0} \\ -\mathbf{R} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ -\mathbf{R} & \mathbf{0} & \mathbf{0} & \mathbf{Q}_d \end{bmatrix} \mathbf{w}_k < 0. \tag{24}$$

Subsequently, by application of a so-called S-procedure [4] it can be shown that a following result is attained

$$\mathbf{w}_k^T \begin{bmatrix} \mathbf{W}_1 & \tilde{\mathbf{X}}_1^T \mathbf{R} \tilde{\mathbf{X}}_2 & \tilde{\mathbf{X}}_1^T \mathbf{R} \tilde{\mathbf{Z}}_1 & \tilde{\mathbf{X}}_1^T \mathbf{R} \tilde{\mathbf{Z}}_2 \\ \tilde{\mathbf{X}}_2^T \mathbf{R} \tilde{\mathbf{X}}_1 & \mathbf{W}_2 & \tilde{\mathbf{X}}_2^T \mathbf{R} \tilde{\mathbf{Z}}_1 & \tilde{\mathbf{X}}_2^T \mathbf{R} \tilde{\mathbf{Z}}_2 \\ \tilde{\mathbf{Z}}_1^T \mathbf{R} \tilde{\mathbf{X}}_1 & \tilde{\mathbf{Z}}_1^T \mathbf{R} \tilde{\mathbf{X}}_2 & \tilde{\mathbf{Z}}_1^T \mathbf{R} \tilde{\mathbf{Z}}_1 - \alpha \mathbf{Q}_d & \tilde{\mathbf{Z}}_1^T \mathbf{R} \tilde{\mathbf{Z}}_2 \\ \tilde{\mathbf{Z}}_2^T \mathbf{R} \tilde{\mathbf{X}}_1 & \tilde{\mathbf{Z}}_2^T \mathbf{R} \tilde{\mathbf{X}}_2 & \tilde{\mathbf{Z}}_2^T \mathbf{R} \tilde{\mathbf{Z}}_1 & \tilde{\mathbf{Z}}_2^T \mathbf{R} \tilde{\mathbf{Z}}_2 - \beta \mathbf{Q}_d \end{bmatrix} \mathbf{w}_k < 0, \tag{25}$$

where $\mathbf{W}_1 = \tilde{\mathbf{X}}_1^T \mathbf{R} \tilde{\mathbf{X}}_1 + \mathbf{S} - (1 - \alpha - \beta) \mathbf{R}$, $\mathbf{W}_2 = \tilde{\mathbf{X}}_2^T \mathbf{R} \tilde{\mathbf{X}}_2 - (1 - \alpha - \beta) \mathbf{S}$. Afterwards, by applying Schur complement to (25) with left- and right-side multiplication by $\text{diag}(\mathbf{I}, \mathbf{I}, \mathbf{I}, \mathbf{I}, \mathbf{R})$, it leads to

$$\begin{bmatrix} \mathbf{S} - (1 - \alpha - \beta) \mathbf{R} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \tilde{\mathbf{X}}_1^T \mathbf{R} \\ \mathbf{0} & -(1 - \alpha - \beta) \mathbf{S} & \mathbf{0} & \mathbf{0} & \tilde{\mathbf{X}}_2^T \mathbf{R} \\ \mathbf{0} & \mathbf{0} & -\alpha \mathbf{Q}_d & \mathbf{0} & \tilde{\mathbf{Z}}_1^T \mathbf{R} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & -\beta \mathbf{Q}_d & \tilde{\mathbf{Z}}_2^T \mathbf{R} \\ \mathbf{R} \tilde{\mathbf{X}}_1 & \mathbf{R} \tilde{\mathbf{X}}_2 & \mathbf{R} \tilde{\mathbf{Z}}_1 & \mathbf{R} \tilde{\mathbf{Z}}_2 & -\mathbf{R} \end{bmatrix} \prec 0, \quad (26)$$

Setting up $\mathbf{R} \tilde{\mathbf{X}}_1 = \mathbf{R} \tilde{\mathbf{A}}_1 - \mathbf{R} \tilde{\mathbf{F}} \tilde{\mathbf{C}} = \mathbf{R} \tilde{\mathbf{A}}_1 - \mathbf{N} \tilde{\mathbf{C}}$, $\mathbf{R} \tilde{\mathbf{X}}_2 = \mathbf{R} \tilde{\mathbf{F}} \tilde{\mathbf{A}}_2 = \mathbf{N} \tilde{\mathbf{A}}_2$, $\mathbf{R} \tilde{\mathbf{Z}}_1 = \mathbf{R} \tilde{\mathbf{D}}_1 - \mathbf{R} \tilde{\mathbf{F}} \tilde{\mathbf{D}}_2 = \mathbf{R} \tilde{\mathbf{D}}_1 - \mathbf{N} \tilde{\mathbf{D}}_2$, $\mathbf{R} \tilde{\mathbf{Z}}_2 = \mathbf{R} \tilde{\mathbf{F}} \tilde{\mathbf{D}}_3 = \mathbf{N} \tilde{\mathbf{D}}_3$, where $\mathbf{N} = \mathbf{R} \tilde{\mathbf{F}}$. Employing these set up into (26) leads to (18) which results that the proof is completed.

Hence, finally the design procedure boils down to solving the set of LMIs (18) and then, taking into account the fact that $\mathbf{N} = \mathbf{R} \tilde{\mathbf{F}}$, to calculate gain matrices for the estimator as $\tilde{\mathbf{F}} = [\mathbf{F}_x^T \ \mathbf{F}_u^T]^T = \mathbf{R}^{-1} \mathbf{N}$.

The aim of the consecutive section is to show the performance of the proposed approach.

4 Illustrative Example

Taking into account a novel approach, provided in the former section and capable of reconstructing not only the state of the system but also the faults influencing onto this system, especially both actuator fault as well as sensor fault, the validation of such an approach has been made by implementing it to the Multi-Tank system provided by Inteco Ltd. [11] shown in Fig 1. For more information the reader is referred to [26].



Fig. 1. The laboratory Multi-tank system

The actuator fault distribution matrix \mathbf{L}_u is given as $\mathbf{L}_u = [1.1429 \ 0 \ 0]^T \neq \mathbf{B}$. To indicate the performance, let the output matrix be given as $\mathbf{C} = \mathbf{I}_{2 \times 2}$ which denotes that the state variable for the liquid level in the lower tank was immeasurable during the experiment to make it more difficult, which entails the sensor fault distribution matrix set to $\mathbf{L}_y = [0 \ 1]^T$.

Such a configuration of the system entails a following fault scenario:

$$f_{u,k} = \begin{cases} -0.45 \cdot u_p & 7000 \leq k \leq 11000 \\ 0 & \text{otherwise} \end{cases} \quad (27)$$

$$f_{y,1,k} = \begin{cases} 0.03 \sin(k \cdot 5 \cdot 10^{-3}) & 6000 \leq k \leq 13000 \\ -0.1 & 13001 \leq k \leq 16500 \\ 0 & \text{otherwise} \end{cases} \quad (28)$$

Such involved fault scenario with pre-defined fault profiles, allows for investigating abrupt loss of efficiency of the pump and then slow fluctuation of its degradation. The other aspect of the fault scenario contribute to evaluating the sensor bias and measurement reading shifting in which the measured value of the liquid level is shifted by some value respectively to the real level in the tank.

Having such defined fault scenario, let us also demonstrate an enhancement of the proposed approach in this paper by comparing it with the approaches allowing for estimating the actuator fault or sensor fault only, introduced in e.g. [20,24]. Such a comparison will allow to answer the questions, what will happen when actuator/sensor fault occurs while estimating the other one kind of fault?

For further analysis it should be pointed out that the system has performed in an open-loop with a predefined control signals set as the ones presented in Fig. 2 during the experiment. From this figure it can be easily noticed that the pump speed as well as solenoid valves were operating in a different manner during the experiment. It entails the constant changing of the state of the entire system.

To compare the proposed approach with the ones allowing for reconstruction of the only one kind of faults, Figs. 3a-c present particular states of the considered system with their estimations. In these Figs., the results achieved with the proposed in this paper algorithm are marked with ASFE which refers to actuator and sensor fault estimator. It should be emphasized that the results obtained with actuator fault estimator (AFE) and sensor fault estimator (SFE) have been obtained by designing the estimators accordingly to approaches proposed in [20,24]. In these diagrams, blue solid lines represent the real state while green dash-dotted lines stand for

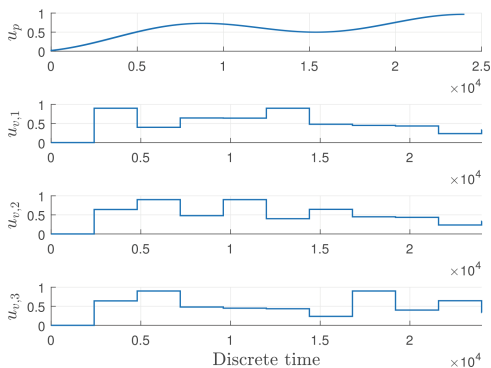
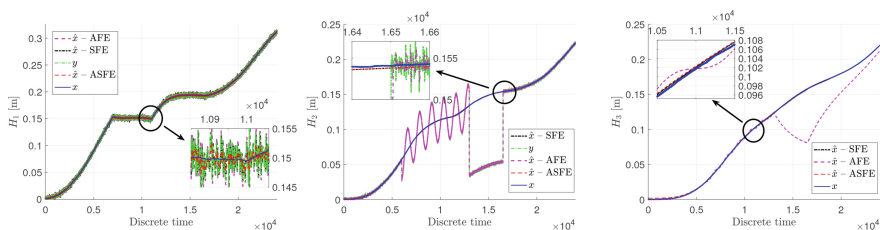


Fig. 2. The control signals u during the experiment.



(a) The liquid level in the top tank (b) The liquid level in the middle tank (c) The liquid level in the lower tank

Fig. 3. The liquid level in the top, middle and lower tanks

the measured output. Moreover, the red dashed lines represent the state estimate obtained with ASFE whilst the dash-dotted black lines along with dashed magenta lines stand for the state estimates achieved with SFE and AFE, respectively. It can be easily observed that each state of the system was appropriately recovered with using ASFE. In the other words, this state estimation follows the real state irrespectively of the actuator fault or the sensor fault. The opposite situation is attained with the others. Despite the actuator fault influencing the pump which caused the water level in all tanks to be disturbed, the very first state was recovered incredibly correctly with all three estimators. However, due to the fact that the sensor fault occurred in the middle tank, the state was estimated correctly only by the ASFE as well as SFE. Indeed, this is not a big astonishment as these estimators were developed to cope with sensor faults in such a way to minimize the error between the real state and estimated one while the sensor fault occur. In such cases the state estimates track the real one despite the fact that sensor readings were given incorrectly. Whereas, the actuator fault has estimated the measurements of the system instead of the state. The third state however, was recovered appropriately despite the fact that it was immeasurable with one exception. The AFE based on its incorrect estimation of the water level in the middle tank, assessed the level in third tank to completely different values.

5 Conclusions

The paper dealt with the problem of actuator and sensor faults estimation for linear dynamic systems. The system was considered under influence of unknown exogenous disturbances. The proposed approach states for the fusion of the solutions available in the literature which results that an estimator without unwelcome rate of change of actuator fault factor is developed. The proposed approach was implemented to the laboratory Multi-Tank system and the results confirms the performance of the approach. In the future work the authors will focus on developing an FTC scheme utilizing the estimation approach proposed in this paper.

References




1. Alessandri, A., Baglietto, M., Battistelli, G.: Design of state estimators for uncertain linear systems using quadratic boundedness. *Automatica* **42**(3), 497–502 (2006)
2. Aouaouda, S., Chadli, M.: Robust fault tolerant controller design for Takagi-Sugeno systems under input saturation. *Int. J. Syst. Sci.* **50**(6), 1163–1178 (2019)
3. Ashfaq, B.S., Tsakalis, K.: Discrete-time PID controller tuning using frequency loop-shaping. *IFAC Proceed. Vol.* **45**(3), 613–618 (2012)
4. Boyd, S., Feron, E., Ghaoui, L.E., Balakrishnan, V.: *Linear Matrix Inequalities in System and Control Theory*. SIAM, Philadelphia (1994)
5. Ding, B.: Constrained robust model predictive control via parameter-dependent dynamic output feedback. *Automatica* **46**(9), 1517–1523 (2010)

6. Ding, B.: Dynamic output feedback predictive control for nonlinear systems represented by a Takagi-Sugeno model. *IEEE Trans. Fuzzy Syst.* **19**(5), 831–843 (2011)
7. Fadhilah, H.N., Adzkiya, D., Arif, D.K., Zhai, G., et al.: Decentralized static output feedback controller design for linear interconnected systems. *Int. J. Appl. Math. Comput. Sci.* **33**(1), 83–96 (2023)
8. Gao, Z., Ding, S.X., Cecati, C.: Real-time fault diagnosis and fault-tolerant control. *IEEE Trans. Industr. Electron.* **62**(6), 3752–3756 (2015)
9. He, W., Meng, T., He, X., Ge, S.S.: Unified iterative learning control for flexible structures with input constraints. *Automatica* **96**, 326–336 (2018)
10. Ijaz, S., Yan, L., Hamayun, M.T., Baig, W.M., Shi, C.: An adaptive LPV integral sliding mode FTC of dissimilar redundant actuation system for civil aircraft. *IEEE Access* **6**, 65960–65973 (2018)
11. INTECO: Multitank System - User's manual. INTECO. <http://www.inteco.com.pl/> (2013)
12. Kantue, P., Pedro, J.O.: Integrated fault-tolerant control of a quadcopter UAV with incipient actuator faults. *Int. J. Appl. Math. Comput. Sci.* **32**(4), 601–617 (2022)
13. Kukurowski, N., Pazera, M., Witczak, M.: Fault-tolerant tracking control for a descriptor system under an unknown input disturbances. *Electronics* **10**(18), 2247 (2021)
14. Lamouchi, R., Raissi, T., Amairi, M., Aoun, M.: Interval observer-based methodology for passive fault tolerant control of linear parameter-varying systems. *Trans. Instit. Measure. Control* **44**, 01423312211040370 (2021)
15. Lan, J., Patton, R.J.: Robust integration of model-based fault estimation and fault-tolerant control. Springer, Cham (2020). <https://doi.org/10.1007/978-3-030-58760-4>
16. Liu, M., Shi, P.: Sensor fault estimation and tolerant control for itô stochastic systems with a descriptor sliding mode approach. *Automatica* **49**(5), 1242–1250 (2013)
17. Lunze, J., Lehmann, D.: A state-feedback approach to event-based control. *Automatica* **46**(1), 211–215 (2010)
18. O'Dwyer, A.: An overview of tuning rules for the pi and PID continuous-time control of time-delayed single-input, single-output (SISO) processes. In: Vilanova, R., Visioli, A. (eds.) *PID Control in the Third Millennium. Advances in Industrial Control*. Springer, London (2012). https://doi.org/10.1007/978-1-4471-2425-2_1
19. Owens, D.H., Feng, K.: Parameter optimization in iterative learning control. *Int. J. Control* **76**(11), 1059–1069 (2003)
20. Pazera, M., Buciakowski, M., Witczak, M.: Robust multiple sensor fault-tolerant control for dynamic non-linear systems: application to the aerodynamical twin-rotor system. *Int. J. Appl. Math. Comput. Sci.* **28**(2), 297–308 (2018)
21. Pazera, M., Witczak, M., Korbicz, J.: Combined estimation of actuator and sensor faults for non-linear dynamic systems. In: 2017 22nd International Conference on Methods and Models in Automation and Robotics (MMAR), pp. 933–938. IEEE (2017)
22. Pazera, M., Witczak, M., Kukurowski, N.: Robust unknown input observer design for simultaneous actuator and sensor faults. In: 2019 24th International Conference on Methods and Models in Automation and Robotics (MMAR), pp. 366–371. IEEE (2019)
23. Pazera, M., Witczak, M., Kukurowski, N., Buciakowski, M.: Towards simultaneous actuator and sensor faults estimation for a class of Takagi-Sugeno fuzzy systems: a twin-rotor system application. *Sensors* **20**(12), 3486 (2020)

24. Rotondo, D., Witczak, M., Puig, V., Nejjari, F., Pazera, M.: Robust unknown input observer for state and fault estimation in discrete-time Takagi-Sugeno systems. *Int. J. Syst. Sci.* **47**(14), 1–16 (2016)
25. Van, M., Ge, S.S., Ren, H.: Robust fault-tolerant control for a class of second-order nonlinear systems using an adaptive third-order sliding mode control. *IEEE Trans. Syst. Man Cybern. Syst.* **47**(2), 221–228 (2016)
26. Witczak, M.: Fault Diagnosis and Fault-Tolerant Control Strategies for Non-Linear Systems. Volume 266 of *Lectures Notes in Electrical Engineering*. Springer International Publisher, Springer, Cham (2014), <https://doi.org/10.1007/978-3-319-03014-2>
27. Witczak, M., Buciakowski, M., Mrugalski, M.: An \mathcal{H}_∞ approach to fault estimation of non-linear systems: Application to one-link manipulator. In: 19th International Conference On Methods and Models in Automation and Robotics (MMAR) 2014, pp. 456–461. IEEE (2014)
28. Yin, S., Luo, H., Ding, S.X.: Real-time implementation of fault-tolerant control systems with performance optimization. *IEEE Trans. Industr. Electron.* **61**(5), 2402–2411 (2013)
29. Zhang, Z., Yang, Z., Liu, S., Chen, S., Zhang, X.: A multi-model based adaptive reconfiguration control scheme for an electro-hydraulic position servo system. *Int. J. Appl. Math. Comput. Sci.* **32**(2), 185–196 (2022)



Enhancing Power Generation Efficiency of Piezoelectric Energy Harvesting Systems: A Performance Analysis

Bartłomiej Ambrozkiewicz¹ (✉) , Zbigniew Czyż² , Paweł Stączek¹ ,
Jakub Anczarski¹, and Mikołaj Jachowicz¹

¹ Lublin University of Technology, Nadbystrzycka 36, 20-618 Lublin, Poland
b.ambrozkiewicz@pollub.pl

² Polish Air Force University, Dywizjonu 303 35, 08-521 Dęblin, Poland

Abstract. This study is aimed to analyze a piezoelectric energy harvesting system made of a smart lead zirconate material applied to the hybrid excitation in the form of mechanical vibrations and airflow. The system utilized a monolithic PZT plate composed of ceramic-based piezoelectric material, under its bending, the voltage is generated. In the experiment, a specialized measurement system based on Arduino and wind tunnel were used to conduct the tests on a bluff-body shape mounted on an elastic beam with a piezoelectric attached to a support with arms. The elastic beam and arms were linked with springs, that change the internal characteristic frequency of the energy harvester. The air-flow velocity and forced vibration frequencies were varied, and the output voltage signal and linear accelerations were recorded during the variable excitation. The effect study is the conclusion that the system's highest voltage efficiency was achieved through a combination of mechanical vibrations and air flow. Additionally, the study generated the correlation between the RMS voltage as a function of the air-flow velocity for resonance excitation frequencies. In the study, the area where the system could generate the highest output voltage under given excitation conditions was investigated. The above-described comparative analysis is conducted for two research objects differing in their mass. The proposed energy harvesting system concept can be used as a power supply for low-power consumption sensors.

Keywords: Hybrid power generation system · macro fiber composite material · structural cross-sectional configuration · vortex-induced vibration (VIV) · aeroelastic instability

1 Introduction

As the world seeks to reduce its dependence on fossil fuels, new sources of green energy are being explored. One promising way is the energy harvesting [1, 2], which is a process that involves scavenging and converting sources such as mechanical vibrations, temperature gradients, or light into small amounts of power that can be used to supply remote devices with low-power consumption. This paper focuses on combining two types

of energy harvesting: vibrational energy harvesters (VEHs) and systems that scavenge energy from airflow, what can be treated as the hybrid system. These systems can be categorized based on their energy conversion method, including electromagnetic [3], piezoelectric [4], and electrostatic effects [5]. Over the past two decades, there has been a continuous development of new designs for energy harvesting systems, with research results compiled in reviews [6, 7]. The latest trend is focused on hybrid energy harvesters that utilize multiple effects or phenomena to scavenge electrical energy, such as following combinations of piezoelectric and triboelectric effects, piezoelectric and electromagnetic effects, or piezoelectric effects and water waves. This paper will specifically discuss a hybrid aeroelastic energy harvesting system that combines piezoelectric effects with air-flow [8, 9].

In the process of design of the energy harvesting system, maximizing power output is a key consideration, what can be for instance achieved by adjusting the system to operate just at or near its resonance frequency, or by utilizing subharmonic solutions during system's operation. The previous research focused on designing a system that is susceptible to both vortex-induced vibrations (VIVs) at low wind speeds and the galloping effect at high wind speeds [10]. Studies on wind energy harvesting systems suggest that circular bluff bodies exhibit oscillatory behavior due to VIVs, while square-shaped bodies generate the highest power output through galloping. To combine both effects and increase power performance across a wider range of wind velocities, we proposed newly designed system that combines circular and square shapes along its generatrix. However, this system operated best at higher wind speeds and was susceptible to the galloping effect. To address this, we propose a modified design with additional springs to increase sensitivity to low wind speeds by changing system's internal characteristic frequency. In the system, 3 properties related to springs can change the frequency, i.e. the mounting position of springs on the beam, the spring's stiffness and the width between beam and arms. In this paper, we focus only on one chosen position for springs on linked to the beam.

To improve the power output of the system, a hybrid excitation approach was adopted, combining both air flow and mechanical vibrations. This approach has been previously studied in following references [11, 12] and has the advantage of creating a desirable dynamical system of an energy harvester by specific excitation conditions from both wind velocity and a mechanical vibrations shaker. Such an approach increases the flexibility of the system to variable excitation conditions and affect the voltage output of the system. The primary objective of the new design, in contrast to the previous one, is to obtain an oscillating solution for low wind velocities. Additional springs were mounted on a beam as a countermeasure to change its characteristic frequency and make it more flexible for a wider range of excitations. Referring to the previous configuration, the oscillating solutions were observed by higher values of air-flow velocity, i.e. $U = 8\text{--}9$ m/s. Introduced modifications in the energy harvester helped to obtain the desired solution in a wide range. The experimental setup involved testing the system under various wind velocities and frequencies of excitation finding the most desirable results from the perspective of output voltage. The results showed that the new design with additional springs improved the power output at low wind velocities, demonstrating the effectiveness of the countermeasures.

The remainder of this paper is following, in Sect. 2, the experimental setup consisting of energy harvesting concept and measurement circuit are presented. In Sect. 3, the results

of the energy harvesting system performance are presented and discussed. Section 4 summarizes the paper defining the future studies on the system.

2 Experimental Setup

The studied energy harvesting system concept consists of 2 main parts, i.e. electronic measurement system in which 2 accelerometers can be distinguished, i.e. accelerometer mounted at the tip mass and accelerometer mounted on the shaker and the elastic beam with piezo-element on which the bluff-body is mounted (Fig. 1). The piezoelectric system comprises an elastic beam made of aluminum, a piezoelectric element, a bluff body, and a spring arm to attach additional springs, the other ends of which are connected to the elastic beam. The total length of the elastic beam is equal to 200 mm, with a cross-sectional width and thickness of 20×1 mm. The section of the elastic beam without the attachment point measures 160 mm, with 20 mm allotted for attachment to the sting and another 20 mm for attachment to the bluff body. The beam was secured to a support structure, which was a vertical aluminum flat bar measuring 20×4 mm in cross-section. The bluff body spring arm was 3D-printed using PLA (Polylactide) material and was bolted to the support structure at the height of the bluff body action. The orientation of the bluff body and elastic beam is such that they allow for horizontal oscillation, which results from the induced vortices behind the bluff body.

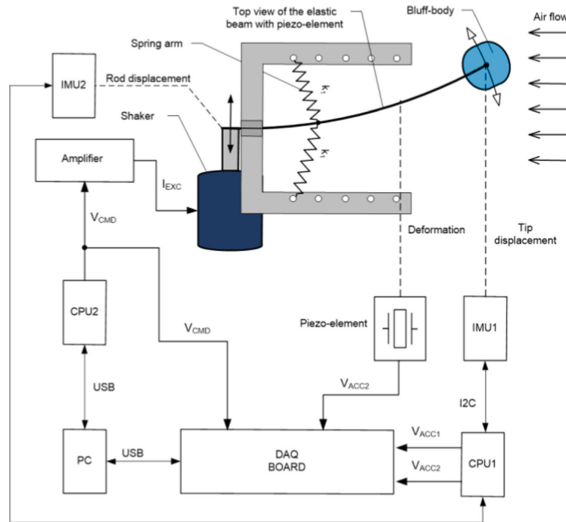


Fig. 1. Measurement circuit used for the data acquisition in the experiment. The bluff-body is presented in the top view. Full details on the elements of the test rig can be find in the reference [13].

The piezoelectric material utilized in the study was a type P1 Macro Fiber Composite™ (MFC) manufactured by Smart Material GmbH (Dresden, Germany). This type of piezoelectric material is composed of a monolithic PZT plate, which is a piezoelectric ceramic, sandwiched between interdigitated electrode patterns on a polyimide film. The piezoelectric element has two distinct areas: an active area of 28×14 mm and an overall dimension of 38×20 mm. The piezoelectric element can operate within a voltage range

of -500 V to $+1500$ V and has a maximum operating frequency of fewer than 1 MHz. The piezoelectric element is $300\ \mu\text{m}$ thick, and its capacitance is $1.9\ \text{nF}$ ($\pm 20\%$), for furthermore extended details on this particular piezoelectric element, refer to [11].

Vortices and corresponding vibrations are induced by the bluff body, which is a combination of a cylinder and a cuboid. The geometry of the bluff body was designed using CAD software and then printed using the Fused Deposition Modeling (FDM) method on a 3D printer. The weight of the prepared bluff bodies used in the experiment were $30.4\ \text{g}$ for 60% infill and $37.8\ \text{g}$ for 80% infill, for the extended experiments also another value of infill were proposed, i.e. 20%, 40% and 100%. The shape of the bluff body's cross-section smoothly transitions from a square to a circle along its entire length, using a spline function. It's important to note that the cross-sectional areas of the square and circle at their respective ends are identical and equal to $400\ \text{mm}^2$, what refers to the constant Reynolds number. The changing cross-section along the generatrix of bluff-body refers to the VIV and galloping phenomena.

The experiment was conducted in an open wind tunnel, model GUNT HM 170, with a closed measurement section. The measurement section of the wind tunnel is a square with dimensions of $300\ \text{mm} \times 300\ \text{mm}$ and the test object was placed at the center of this section. The air flow velocity around the test object was controlled by adjusting the speed of a fan located at the outlet of the wind tunnel. Figure 2 illustrates the test object in the measurement section of the wind tunnel. A TIRA S513 vibration generator was used to excite the elastic beam and it was positioned underneath the measurement section.



Fig. 2. The measurement section of the wind tunnel includes several components including: 1 - Bluff-body made in FDM printing technique, 2 - 3-axis MEMS digital accelerometer module at the tip of the beam, 3 - two Arduino microcontroller boards, 4 - National Instruments Data Acquisition Card, 5 - Accelerometer module mounted on the shaker plate, 6 - Piezo-patch. Full details on the elements of the test rig can be find in the reference [13].

Details of the wind tunnel can be found in the following reference [10]. This device is capable of producing a rated peak force of 100 N, and its frequency ranges from 2 to 7000 Hz, enabling an axial displacement of 13 mm, a maximum velocity of 1.5 m/s, and a maximum acceleration of 45 g. The sinusoidal excitation signal was amplified during the measurements using a TIRA DA 200 digital amplifier.

3 Results and Discussions

For the recognition of the system's dynamics, the test was conducted for 2 different bluff-bodies differing with the infill, i.e. 60% and 80%. In the first stage of the experiment, only the mechanical vibrations were used as the excitation in the wide range, i.e. from 0.5 Hz to 5.0 Hz with step of 0.5 Hz. However, the shorter step of excitation frequency was applied around found internal resonance of each mass. As a result, for the mass characterizing 60% infill, its internal resonance was found for the excitation of $f = 2.75$ Hz and for the 80% infill around $f = 2.40$ Hz. The resonance curves presented in Fig. 3 were obtained from the acceleration signal of the tip mass, and next its RMS value was calculated from mentioned time-series. In Table 1, the value of excitations is specified both for airflow velocity and frequency of excitation (by found characteristic frequency). For above found resonances for each mass, the experiment was conducted, for the variable value of the airflow velocity. Due to the reduced stiffness of the bluff-body support, there is a little shift in the frequency of excitation and the frequency at the tip-mass [13], nevertheless, the structure is stable enough to provide realistic results. For each mass, what is worth to observe in Fig. 3 is the value of amplitude, for lighter mass, the acceleration is higher and the mass is moving freely. On the other hand for the second mass of bluff body with infill 80%, the mounted springs are having stronger impact on the oscillating system.

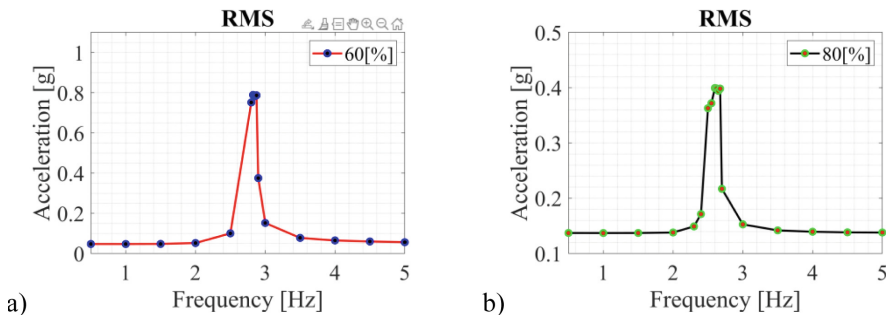


Fig. 3. The resonance curves obtained for following masses of bluff-body with infill a) 60% and b) 80%. The step for considered cases of frequency points are close to each other by found resonance and oscillating solution of the bluff-body.

Next step of the experiment is the analysis of voltage time-series obtained from the piezo-element. In Fig. 4 the recorded output voltage from the piezoelectric element for the resonant frequencies is presented in the domain of air-flow velocity. As it was mentioned, the results for each mass were obtained for its characteristic frequency (Table 1).

Table 1. Model properties applied for simulation study.

Airflow velocity U [m/s]	Frequency of excitation f [Hz]
4.2–10	2.75–60 [%] 2.40–80 [%]

Referring to the previous configuration without the additional springs [10], higher oscillations are obtained for small values of air-flow velocity which is desirable from the perspective of energy harvesting and low input power. When comparing the results for two weights (see orange and blue points and trend lines in Fig. 4) there is a difference in the peaks of output voltage. These peaks are not only dependent on the respective masses/weights but also the specific value of air-flow velocity. Additionally, the results of output voltage between masses have a similar tendencies within the increase of the air-flow velocity, the RMS value of voltage is decreasing. It refers to the fact, that the higher velocity impact negatively to the dynamics of the system resulting in a lower value of the voltage. The comparative analysis of two case-studies is only a small part of the analysis, while we see the other features of the system, which can impact the dynamics and the voltage/power output of the prototype and they are the following:

- Mass of the bluff-body,
- Width of the frame,
- Mounting position of springs,
- Frequency and amplitude of excitation,
- Air-flow velocity,
- Shape of the bluff-body,
- Application of vibroisolators on frame's sides for adding the impact phenomena to the system,
- Application of additional magnets to change the energy potential in the system.

Overall, the results of the experiments show that the highest output voltage and RMS acceleration can be obtained at a specific combination of wind velocity and excitation frequency. The internal resonance frequency of the system can be easily adjusted by changing the position of additional springs that are coupled with the oscillating beam. These results are important for optimizing the design of energy harvesting systems based on bluff-body oscillations in wind tunnels, and can lead to improved performance and efficiency. In the next part of the experiment, the analysis will be referred to above pointed features, which can be changed in the system. Especially, for the same design, the different position of springs will be considered and the width of arms' position. For the newly prepared concept, the vibroisolators will be used to improve the power efficiency of the system from the analysis of super- or sub-harmonic solutions.

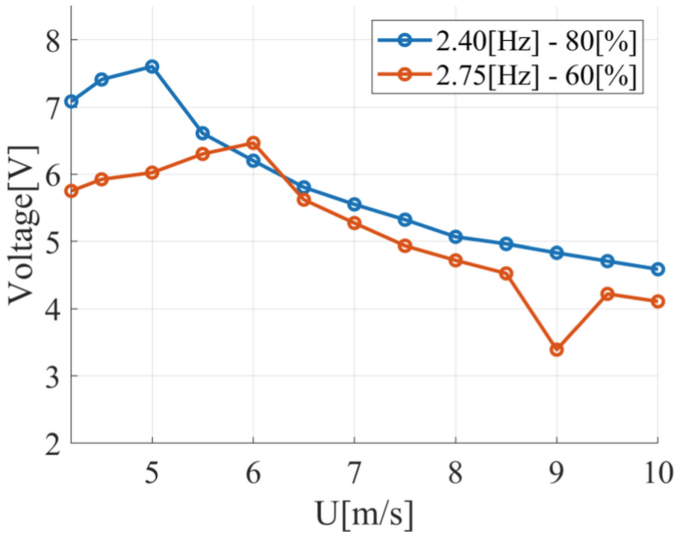


Fig. 4. The RMS voltage on the piezoelectric element as a function of air flow velocity for two research object with different mass, 60% of infill - red, 80% of infill - blue.

4 Summary

The given experimental results show that the bluff bodies vibrate over a wide range of air flow velocities in their characteristic frequency. For the first step, their internal resonance have been found in the experimental way. The dynamical behavior changes with the value of air-flow velocity and the most optimal behavior from the perspective of energy harvesting was observed for small values of air-flow velocities. This fact is positive, that the small value of input energy is needed to obtain the high voltage/power efficiency from the system. This phenomena have been obtained due to the installation of additional springs on sides of the of the elastic beam.

To optimize the system's performance further, side springs were added to achieve an oscillatory solution at both high and low wind speeds. The impact of additional springs on system performance will be investigated in future research, with a focus on maximizing the output voltage/power and how it changes with the internal frequency of the system. The internal frequency will vary with different masses and the alignment of the additional springs. Additionally, the other features can be considered to manipulate the prototype's energetic efficiency.

The system has potential applications in various fields. For example, it can be coupled with rotating blades [14], used for structural health monitoring of constructions [15], or it can be coupled with unmanned aerial vehicles (UAVs) to power low-demanding sensors [16]. Further research is required to explore these potential applications fully. For this case, earlier mentioned features that can be changed in the system, will be considered.

Acknowledgements. The research was funded by National Science Centre, Poland under the project SHENG-2, No. 2021/40/Q/ST8/00362.

References

1. Beeby, S.P., Tudor, M.J., White, N.M.: Energy harvesting vibration sources for microsystems applications. *Meas. Sci. Technol.* **17**, R175–R195 (2006)
2. Wang, Y., Song, Y., Xia, Y.: Electrochemical capacitors: mechanism, materials, systems, characterization and applications. *Chem. Soc. Rev.* **45**, 5925–5950 (2016)
3. Miao, G., Fang, S., Wang, S., Zhou, S.: A low-frequency rotational electromagnetic energy harvester using a magnetic plucking mechanism. *Appl. Energy* **305**, 117838 (2022)
4. Zhang, L., Zhang, F., Qin, Z., Han, Q., Wang, T., Chu, F.: Piezoelectric energy harvester for rolling bearings with capability of self-powered condition monitoring. *Energy* **238**, 121770 (2022)
5. Ren, Z., Wu, L., Pang, Y., Zhang, W., Yang, R.: Strategies for effectively harvesting wind energy based on triboelectric nanogenerators. *Nano Energy* **100**, 107522 (2022)
6. Fang, S., Zhou, S., Yurchenko, D., Yang, T., Liao, W.-H.: Multistability phenomenon in signal processing, energy harvesting, composite structures, and metamaterials: a review. *Mech. Syst. Sig. Process.* **166**, 108419 (2022)
7. Sharma, S., Kiran, R., Azad, P., Vaish, R.: A review of piezoelectric energy harvesting tiles: available designs and future perspective. *Energy Convers. Manag.* **254**, 115272 (2022)
8. Yan, Z., Shi, G., Zhou, J., Wang, L., Zuo, L., Tan, T.: Wind piezoelectric energy harvesting enhanced by elastic-interfered wake-induced vibration. *Energy Convers. Manag.* **249**, 114820 (2021)
9. Tian, H., Shan, X., Cao, H., Xie, T.: Enhanced performance of airfoil-based piezoaeroelastic energy harvester: numerical simulation and experimental verification. *Mech. Syst. Sig. Process.* **162**, 108065 (2022)
10. Ambrozkiewicz, B., Czyż, Z., Karpiński, P., Stączek, P., Litak, G., Grabowski, Ł.: Ceramic-based piezoelectric material for energy harvesting using hybrid excitation. *Materials* **14**, 5816 (2021)
11. Bolat, F.C., Basaran, S., Abdelkefi, A., Wang, J.: Experimental comparative analysis of hybrid energy harvesters exposed to flow-induced vibrations. *Proc. Inst. Mech. Eng. C J. Mech. Eng. Sci.* **237**, 664–672 (2023)
12. Javed, U., Abdelkefi, A.: Characteristics and comparative analysis of piezoelectric-electromagnetic energy harvesters from vortex-induced oscillations. *Nonlinear Dyn.* **95**(4), 3309–3333 (2019). <https://doi.org/10.1007/s11071-018-04757-x>
13. Ambrozkiewicz, B., Czyż, Z., Stączek, P., Tiseira, A.O., Garcia-Tiscar, J.: Performance analysis of piezoelectric energy harvesting system. *Adv. Sci. Technol. Res. J.* **16**, 179–185 (2022)
14. Farias, W.P., Souto, C.R., Castro, A.C.: Design and experimental study of a rotational piezoelectric energy harvester. *J. Instrum.* **17**, P10017 (2022)
15. Gul, W., Tahir, S., Razaq, S.: Modeling simulation and characterization of hybrid energy harvester for structural health monitoring. **25**, 215–221 (2022)
16. Ambroziak, L., Ołdziej, D., Koszewnik, A.: Multirotor motor failure detection with piezo sensor. *Sensors* **23**, 1048 (2023)



Power Quality Issues of Photovoltaic Stations in Electric Grids and Control of Main Parameters Electromagnetic Compatibility

Yaroslav Batsala[✉] and Ivan Hlad

Ivano-Frankivsk National Technical University of Oil and Gas, 15 Karpatska Street,
Ivano-Frankivsk 76019, Ukraine
batsala2012@gmail.com

Abstract. This article presents the results of research on the influence of photovoltaic stations on the quality of electricity in low-voltage networks. The use of hybrid electric networks with combination of AC and DC is proposed. We proposed two main stages of the survey before changing the configuration of the electricity network: Forecasting electricity generation by a photovoltaic power station with monitoring of energy parameters and using power quality analyzers to control several indicators. On the basis of the obtained researches it is proposed to make the analysis on indicators total harmonious distortion and reactive capacity.

Keywords: photovoltaic station · forecasting · power quality analyzers · total harmonious distortion · reactive capacity · lab View

1 Introduction

In Ukraine, demand for electricity has intensified due to Russian invasion and missile strikes. The Ukrainian power system is undergoing a difficult period of tests. Destroyed thermal power plants, as well as distribution devices, lead to consideration of the transition to a more decentralized network, which will consist of a large number of local wind and photovoltaic stations. These changes in the configuration of the electricity network lead to analysis of models of forecasting changes of weather parameters, improvement of network infrastructure, conducting researches of possible problems in integration of sources of renewable energy and their connection to the network.

The first step is to improve the methods of forecasting the level of generation of photovoltaic power plants and optimal management of power grid modes [1]. Forecasts of photovoltaic power in different time and space horizons will help owners of photovoltaic stations and operators of power systems to increase efficiency of the energy market.

The power of the photovoltaic station depends on solar insolation, air temperature, cloudy weather (clear sky, fast instant change of cloudy weather, cloudy day), the angle

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

of fall of sunlight, and flooding. Photovoltaic power plants must have a large supply capacity to enable power system operators to make changes in their generation to balance the power system. The operators can control the capacity of the photovoltaic stations for balancing with rechargeable systems (hydro storage stations, Tesla Powerpack, demand side Management) or by removing part of the photovoltaic station if necessary.

Through open platforms for monitoring electricity generation by photovoltaic power stations, scientists receive a set of statistical data. For example, the Solcast platform will help you get free access to direct data and forecasts for private photovoltaic stations (API World Solar Radiation, World PV Power API (obtain forecasts and approximate actual data using latitude and longitude), provides free access in the form of a program code, which can be used in other programs as a program code [2] (Fig. 1).

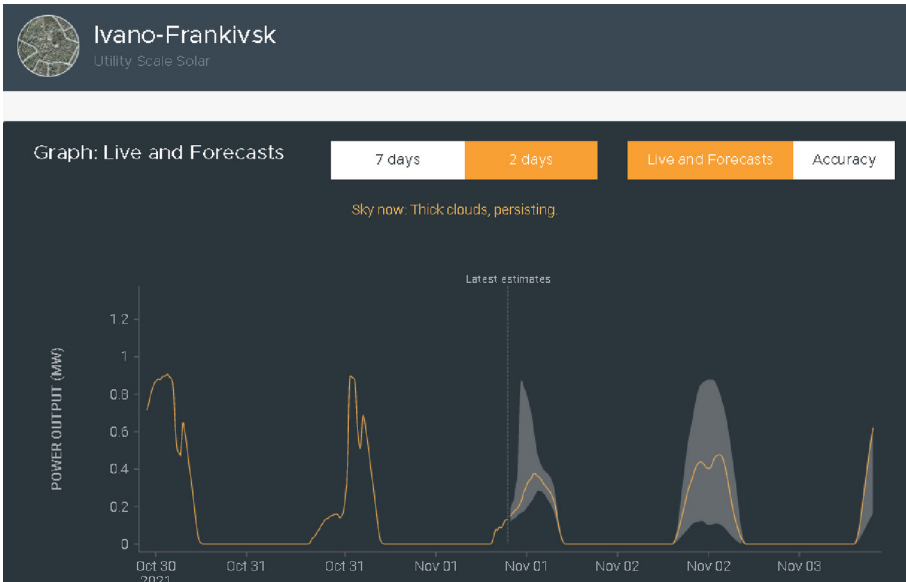


Fig. 1. Forecast data of solar insolation is given for 7 days in Solcast API Toolkit

Most of these platforms provide data on generation for different days, months, as well as changes in solar insolation, voltage and electricity consumption for autonomous stations.

The second step is to analyze the operation of micro-networks with photovoltaic power stations with the help of power quality analyzers and monitoring systems.

2 Literature Review

The latest methods of forecasting electricity generation by photovoltaic power stations and aspects of monitoring of power quality parameters are presented in publications [5–7]. Artificial neural networks can be successfully used for forecasting electricity

generation by renewable power plants, which depend on meteorological parameters. The input data of the ANFIS model – current solar radiation (W/m^2), measured at the local weather station, and the output data – the power of the solar panel (kW). These forecast data may provide more options for analyzing the operation of a photovoltaic station, combining monitoring and modeling, and have different sensitivity and error.

3 Problem Formulation

Application of mathematical models or programs for forecasting will provide information on the level of electricity generation by photovoltaic power stations, which will allow to set up the work of the energy system in crisis moments. It is important to choose the basic parameters of electricity quality analysis for efficient operation of the system as a whole.

4 Results and Discussions

The analysis of the influence of the joint operation of photovoltaic power plants in the electricity network on the quality of electricity confirms the results of research of electromagnetic compatibility indicators (THD - total coefficient of harmonious distortion), reactive power of distortion (T), pulsation coefficient and dose of flicker). With the help of the energy survey the fluctuation of electrical parameters was detected because of non-linear nature of loading in local electrical engineering complexes, which change in time.

The study uses an electricity quality analyzer, which is connected to the bus of the transformer substation, which receives energy from photomodels and converters with the help of current and voltage sensors. Modern power quality analyzers have different functions and costs, can be stationary and portable, provide information online and record monitoring data and a server.

For greater mobility, we used a portable energy quality analyzer that uses current sensors (a set of Rogowski i3000 Flex coils (Fluke firm)) and voltage sensors (LEM CV3-1000), as well as appropriate hardware and software developed in the design environment of LabVIEW virtual appliances. This gives an opportunity to get a long registration of instantaneous values of voltage and currents of the three-phase four-conductor electric network, as well as calculation of average voltage and currents, determination of capacities and harmonious components. As a result, we get the instantaneous values of voltage and current files and the file in the MS Excel environment.

Figure 2 shows the connection of the analyzer to the low voltage bus of the transformer substation.



Fig. 2. Connection of portable quality analyzer to the bus of 0.38 kV transformer substation FES

For analysis and control, we use the built-in tools and functions of LabVIEW, for example, for measuring instantaneous values of currents and voltages, calculating indicators of electromagnetic compatibility (non-sinusoidal voltage coefficients, coefficients of voltage asymmetry in reverse and zero sequence).

Figure 3 shows a fragment of the block diagram in the LabVIEW environment.

For the analysis the dach single-phase photovoltaic station in Ivano-Frankivsk with a capacity of 2.5 kW was chosen. With the use of the power quality analyzer in the sunny day of September at 12.00, the power parameters of the local electricity network were investigated in three modes: 1. The photovoltaic station generates electricity in phase “B”, where there is no consumption; 2. Photovoltaic station does not work; 3 photovoltaic station generates electricity in phase “C” with capacity load, where the amount of consumption is equal to the quantity of generating.

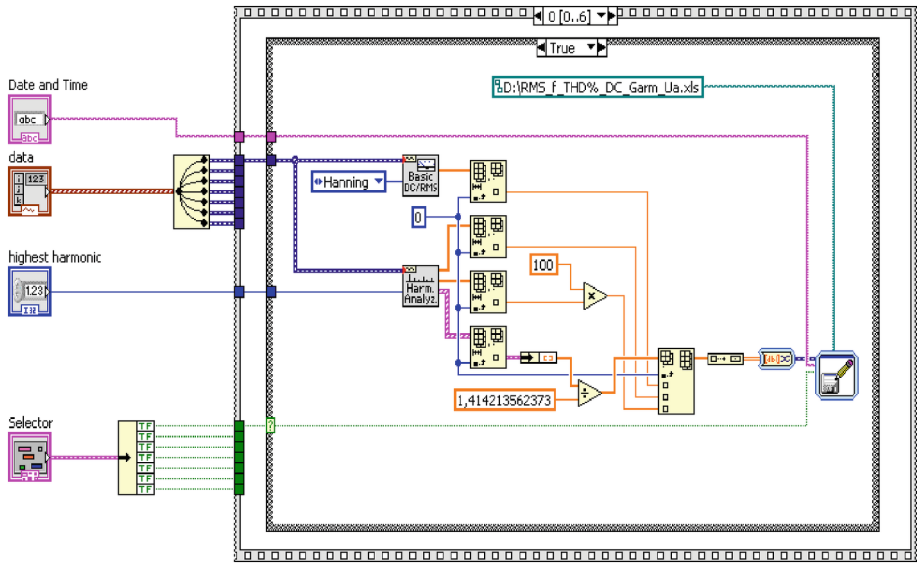


Fig. 3. Fragment of the block diagram in the LabVIEW environment

Figure 4 shows the schedule of change of the active capacity in time in three phases of the enterprise and total capacity (black color).

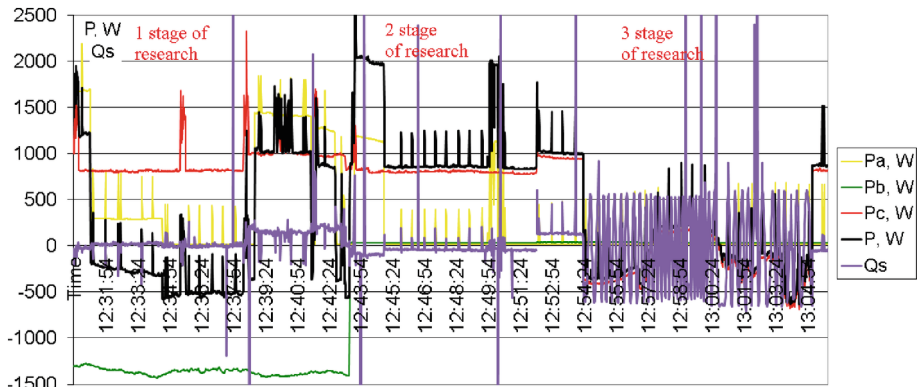


Fig. 4. Graphs of change generated active and reactive power electricity of PVS

The main reasons for the reduction of the quality of electricity and the emergence of voltage and current harmony can be created by non-linear loading through electronic devices, which absorb current high-frequency components. These harmonious components should be reduced directly at the load. Detection of the so-called “resonance zones” will improve the quality of electricity when the output power of the camera-electric stations is equal to the load capacity of the network with a capacious component (cable line, electronic devices) [3] (Fig. 5).

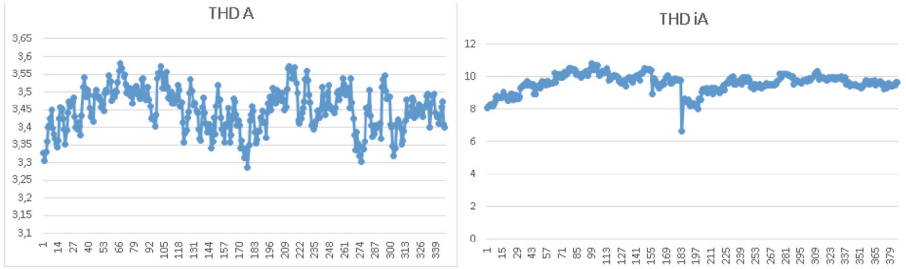


Fig. 5. Graphs of change level THDi and THDu of PVS

Electromagnetic compatibility should be monitored by THD I and reactive power. The research confirms the fact that in days with clear sky of the Internet of 95% of the produced energy put on the network with the value THDI below 5%. In the winter months and the rainy weather, on the contrary, a high percentage of this energy is introduced into the network with THD I values above 5%.

The lack of control over the change in reactive power in power networks with photovoltaic plants can lead to unpredictable consequences, for example, capacitor failures. Transient processes can lead to the failure of sensitive electronic equipment or minimize the duration of operation of network elements, if the amplitude of the transient signal exceeds the permissible limits. To prevent the recommended emergency modes, conduct experimental studies of the operation of solar power plants with simulation of changes in the nature of the loading network and installation of special filters.

The perspective of further research is to expand the functions of the universal hardware and software complex depending on the tasks.

Also, special attention should be paid to solving the issue of accounting for the generation of active energy in the network by one phase. In addition to the phenomenon of the asymmetry of currents and voltages, which also negatively affects the operation of the network, some enterprises have problems with accounting for the generated energy (Fig. 6).

In our opinion, the possibility of a hybrid low-voltage network with photovoltaic power stations, where some of the electricity will be spent by “consumers on a constant stream”, for example, LED lighting, charging phones, fans, laptops, charging batteries of electric locomotives. This will reduce losses in the vectors and reduce harmonious components.

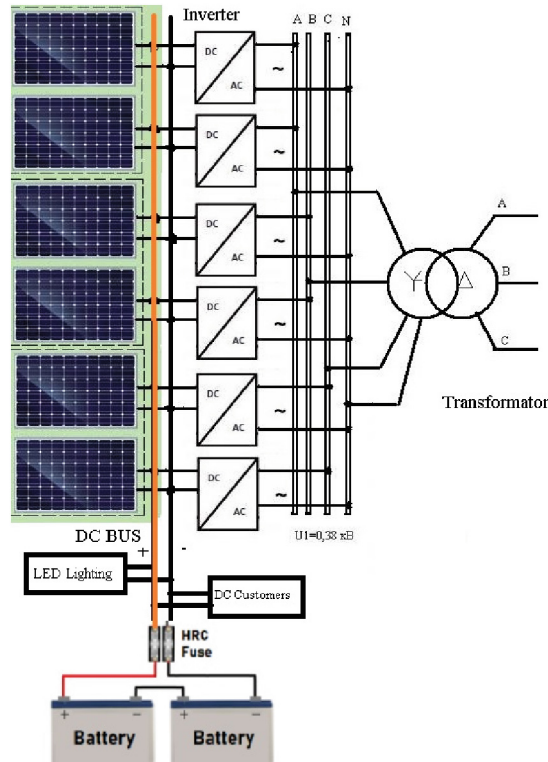


Fig. 6. Hybrid electricity network with constant and alternating current, photo power station and batteries

5 Conclusions

This article analyzes the the need to using power quality analyzers and monitoring systems to detect resonant processes and fluctuations of reactive power, which negatively affects the work of electrical equipment. To prevent emergency modes we recommend conducting experimental research of operation of photovoltaic stations with simulation of change of the nature of network load, weather conditions and installation of special filters.

References

1. Batsala, Ya.V., Hlad, I.V., Yaremak, I.I., Kiiianiuk, O.I.: Mathematical model for forecasting the process of electric power generation by photoelectric stations *Naukovyi Visnyk Natsionalnoho Hirnychoho Universytetu* (1), 111–116 (2021). <https://doi.org/10.33271/nvngu/2021-1/111>
2. Solcast, 2022. Global solar irradiance data and PV system power output data. URL <https://solcast.com/>
3. Hlad, I.V., Batsala, Ya.V.: Influence of solar power plants on low-voltage distribution networks. *Sci. J. "Energy: Econ. Technol. Ecol."* (3 (49)), 82–86 (2017)

4. Hlad, I.V., Batsala, Ya.V.: Experimental tests asymmetric mode in the single-phase network of low power generation solar power. *Oil and Gas Power Eng.* (1), 123–131 (2017)
5. Boris, B., Volodymyr, K., Maryna, N.: The intensity of solar radiation forecasting based on artificial neural networks. In: V International Scientific-Technical Conference Actual Problems of Renewable Energy, Construction and Environmental Engineering, 3–5 June 2021, Kielce, Poland, pp. 19–21 (2021)
6. Kuievda, I., Baliuta, S., Zinkevich, P., Stoliarov, O.: Forecasting the electricity generation of photovoltaic plants. In: V International Scientific-Technical Conference Actual Problems of Renewable Energy, Construction and Environmental Engineering, 3–5 June 2021, Kielce, Poland, pp. 37–38 (2021)
7. Ahsan, S.M., Khan, H.A., Hussain, A., Tariq, S., Zaffar, N.A.: Harmonic analysis of grid-connected solar PV systems with nonlinear household loads in low-voltage distribution networks. *Sustainability* **13**, 3709 (2021). <https://doi.org/10.3390/su13073709>
8. Batsala, Y., Hlad, I., Yaremak, I.: Forecasting day-ahead of power generation from photovoltaic stations and use weather apps. *J. New Technol. Environ. Sci.* (4), 143–149. <https://doi.org/10.53412/jntes-2021-4-3>



Integration of Fault-Tolerant Design and Fault-Tolerant Control of Automated Guided Vehicles

Ralf Stetter^{1(✉)} and Marcin Witczak²

¹ Faculty of Mechanical Engineering, Ravensburg-Weingarten University (RWU), Weingarten, Germany and Steinbeis Transfer Center Automotive Systems, Ravensburg, Germany

stetter@rwu.de

² Institute of Control and Computation Engineering, University of Zielona Gora, 65-246 Zielona Gora, Poland
m.witczak@issi.uz.zgora.pl

Abstract. Both fault-tolerant design (FTD) and fault-tolerant control (FTC) are receiving increasing attention from the scientific community. Both intend to develop and implement solutions for accommodating faults which are inevitable in complex technical systems. However, up to now, little scientific activity was aimed at integrating those two promising approaches. This paper describes a detailed investigation of common aspects and interfaces between FTD and FTC as well as a sensible combined process. This investigation was based on the development of automated guided vehicles (AGVs) together with the appropriate control and diagnosis algorithms and systems.

1 Introduction

In the last years it became apparent that complex technical systems may only function effectively, efficiently and safely, if some kind of fault-tolerant control (FTC) [3, 26] supports the operation and if the system design was consciously carried out and supported by concepts such as fault-tolerant design (FTD) [4, 20]. Under the notion "FTC" algorithms, methods and systems are summarized, which aim at passively or actively accommodating the consequences of faults. Algorithms, methods and tools which aim at supporting engineers to design technical systems which either enable or ease FTC or are fault-tolerant because of certain design characteristics (such as redundancy) are summarized under the notion "FTD". Current research aims at reliability-aware zonotopic tube-based model predictive control [10], graph theory-based approaches [11], parameter identifiability for nonlinear LPV models [19], as well as algorithms and methods for FTD [22, 23]. Despite the prominence of research in both area, the integration of FTD and FTC was not yet in the focus of extensive research activities. These considerations lead to the following central research question: How can FTD and FTC be integrated for more efficient and effective fault accommodation and

higher robustness of technical systems. The structure of the paper was chosen according to this main research question. Section 2 gives background information by describing possible use cases of AGVs and possible occurring faults. In Sect. 3 the integration of FTD and FTC is discussed in detail. An application example - a virtual actuator - is explained in Sect. 4 and Sect. 5 concludes the article.

2 Background

Today, autonomous guided vehicles (AGVs) are applied in nearly all industries because of their efficiency and versatility. Figure 1 shows different AGVs that were developed and realized at the Ravensburg-Weingarten University (RWU).

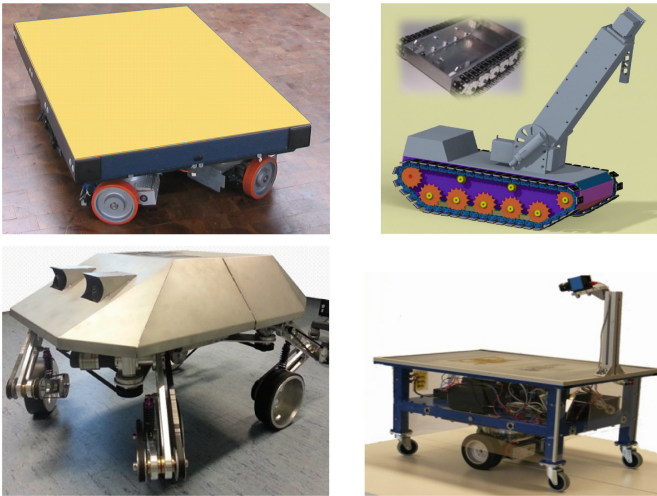


Fig. 1. Different AGVs developed and realized at RWU.

AGVs are for instance used for transporting items inside warehouses or between production stations and can perform both horizontal and vertical transport tasks. AGVs consist of a usually wheel-based load carrier and further components such as robot arms. More and more, AGVs are also applied for the movement of manipulators such as industrial robots which can perform their task at different positions within a manufacturing building. Some AGVs are used to pull or push trailers and are acting as towing vehicle. Another application area is the inspection of technical systems (e.g. pipelines), i.e. AGVs carry some kind of sensor e.g. cameras in order to move along a technical system and be able to inspect different locations. In the different applications, several faults can occur. Figure 2 gives an overview of concerned entities and life-cycle phases in which faults can occur.

	early operation phase	main operation phase	wear & aging phase
sensor fault	sensor fault, early phase	sensor fault, main phase	sensor fault, late phase
actuator fault	actuator fault, early phase	actuator fault, main phase	actuator fault, late phase
process fault	process fault, early phase	process fault, main phase	process fault, late phase
environmental fault	env. fault, early phase	env. fault, main phase	env. fault, late phase
user fault	user fault, early phase	user fault, main phase	user fault, late phase

Fig. 2. Sources of faults and phases within which faults occur.

Several cases are reported concerning faults in sensing elements [1,9] such as faults of sensors in electrical machines or unmanned aerial vehicles. Sensor faults are a common type of fault which are characterized by the phenomenon that a sensing element fails to accurately measure a physical process parameter. This kind of phenomenon can be caused for instance by wiring or calibration problems, vibration, corrosion/oxidation and contamination. Sometimes no sensor signal at all will result from a fault, but also offset sensor output is possible. Possible and probable are also actuator faults, i.e. the undesired condition that an actuator fails to perform its intended function. Common causes for this condition include leakages (for hydraulic actuators), mechanical wear, aging, wiring problems as well as unexpected effects such as buckling. For actuator faults, also saturation is a common fault cause. For AGV systems, actuator faults can be also delays [12]. Faults can also influence the process itself [25]; they occur when the process being does not behave as expected. For AGVs, process faults can be resulting from slippery surfaces. In this case, such faults can also be understood as environmental faults, as the operation environment is deviating from the nominal, specified conditions. Another important source of faults are user faults which can be, for instance, caused by human error, unintended use or even sabotage [2]. Concerning the life-cycle three phases can be distinguished [20]. In the early phase, deficiencies in component production and assembly can be the main causes of faults. In the main operation phase, the predominate faults are stochastic faults. In the late phase of the operation of a technical system additional faults can be caused by aging and wear of components. One example in terms of an AGV is the rechargeable battery. During the life of an AGV and its operation, the useful capacity of a battery and consequently its state of health (SOH) will decrease because of electro-chemical processes [13]. In general, for identifying probable and possible faults, it is sensible to analyse the flows of operands such as matter, energy and signal (compare also [6]) through a technical system - in the given case the AGV.

3 Integration of FTD and FTC

For the integration of FTD and FTC, an initial important insight is that the distinction between active and passive FTC is very important. Passive FTC does

not require fault diagnosis entities and does not rely on sensory capabilities. Still, due to the much larger application potential of active FTC [26], a focus on active FTC is sensible. The main elements of FTC are shown in Fig. 3; certain entities are labelled for the subsequent discussion of the integration of FTD and FTC.

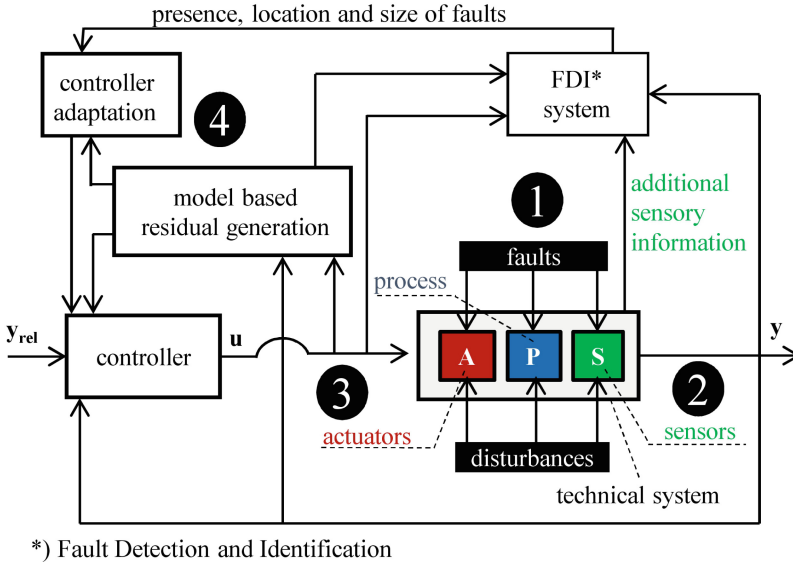


Fig. 3. Active fault-tolerant control (based on [3, 26]).

It is important to note that additional information can be sent from the plant to the fault detection and identification (FDI) system. This information can be data from sensors which do not measure the system output, but e.g. vibrations. It can also be data from actuators that also can generate sensory information e.g. by means of motor current signal analysis (MCSA) [27].

In most systems engineering and design systematics contexts, a distinction of the different levels of abstraction of systems models is proposed [5, 14]. The research of the applicant underlines the necessity to structure the methodical support of FTD according to the different levels; this model of abstraction levels is also sensible for the integration of FTD and FTC. This distinction is, amongst others, shown in Fig. 4.

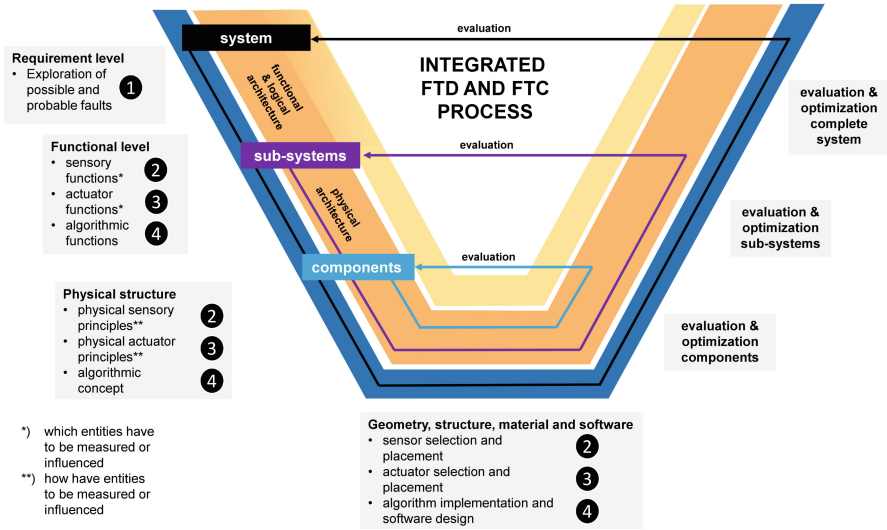


Fig. 4. Integration of FTD and FTC.

Visible in Fig. 4 are the different levels of abstraction requirements level, functional level, level of physical structure and level of geometry, structure, material and software. Visible in Fig. 4 are also sensible locations for the integration of FTC entities in the FTD process structure. An important intention is a conscious and systematic support of an early consideration of FTC potential and necessities. Concrete examples of FTC implementations in FTD processes are given in the subsequent sections.

4 Example: Virtual Actuator

This section describes a concrete integration example - the realization of a virtual actuator. The focus is on the design of a fuzzy virtual actuator (FVA) for AGVs, i.e. on the development of a system structure and algorithms for a control and diagnosis subsystem that operates as FVA. The FVA is combining information of residuals created with an analytical AGV model of the AGV with expert's knowledge. The novel FVA concept is capable to combine information from more than one residual in order to enlarge the decision certainty in fault detection and isolation [21]. In literature, different forms of virtual actuators are described [15–18], however, only one class of fuzzy virtual actuators is mentioned; this class is employing representations of nonlinear systems through Takagi-Sugeno fuzzy models. The demonstration of the feasibility of the novel FVA was possible based on an AGV for rescue missions. For this AGV, the analysis of the forces and torques lead to the formulation of a discretized state space model:

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{E}_k \mathbf{d}_k + \mathbf{W} \mathbf{w}_k. \quad (1)$$

with

$$\mathbf{E}_k = T_s \cdot \mathbf{E}, \quad (2)$$

in this equation \mathbf{w}_k denotes an exogenous disturbance vector with a known distribution matrix \mathbf{W} .

An initial stage in the application of a FVA is the detection of faults. One needs to note that the AGV behavior formulated in Eq. 1 has an unknown input d_k . It is necessary to estimate this input for fault detection and identification. In earlier research, adaptive estimators were proposed for estimating unknown inputs [24]. For the given vehicle, the measurement can be described using the following equation:

$$\mathbf{y}_k = \mathbf{C}_k \mathbf{x}_k + \mathbf{V} \mathbf{v}_k \tag{3}$$

where \mathbf{v}_k denotes the measurement noise and \mathbf{V} stands for the known distribution matrix, which is assumed to be known. For the given FVA an unknown input estimator is employed which is based on one proposed earlier in literature [7] and can be formulated in the following form:

$$\hat{d}_{k-1} = \mathbf{M}_k (\mathbf{y}_k - \mathbf{C}_k \hat{\mathbf{x}}_{k-1|k-1}). \tag{4}$$

Here, $\hat{\mathbf{x}}_{k-1|k-1}$ denotes an estimate of \mathbf{x}_{k-1} . It is possible to define the innovation $\tilde{\mathbf{y}}_k$ by:

$$\tilde{\mathbf{y}}_k \triangleq \mathbf{y}_k - \mathbf{C}_k \hat{\mathbf{x}}_{k-1|k-1}. \tag{5}$$

This underlying estimation process is realized by estimating the unknown input \hat{d}_{k-1} from the measurement \mathbf{y}_k employing the matrix \mathbf{M}_k . The evaluation was based on an assumed fault scenario consisting of an environmental fault - a reduced friction. The simulated set of forces and total torque is depicted in Fig. 5.

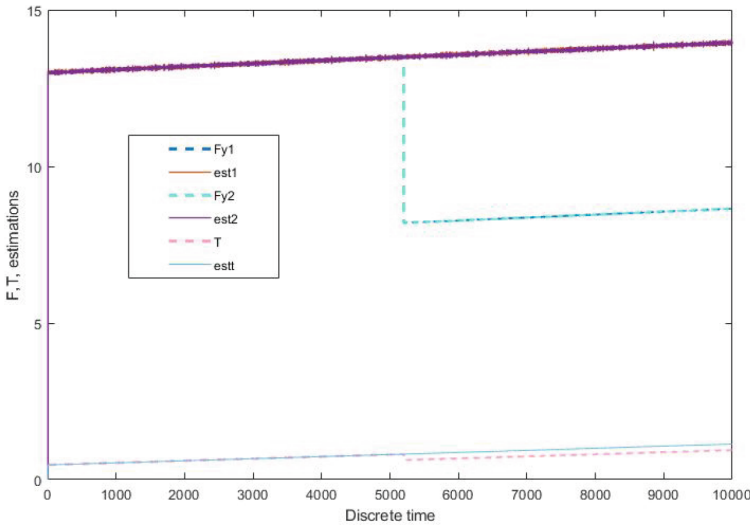


Fig. 5. Fault in Estimations.

From this estimation and sensor readings, residuals can be generated, which are the input of the FVA. The structure of the FVA is shown in Fig. 6.

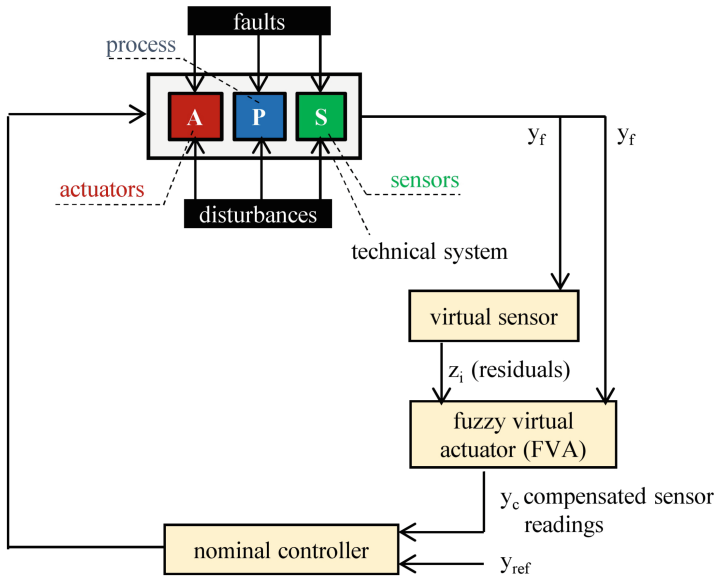


Fig. 6. Structure of a fuzzy virtual actuator.

For the respective AGV, the expert’s knowledge concerning a specific fault could be captured employing three membership functions for each of the residuals z_i which serve as inputs of the fuzzy inference system (FIS). The first input membership function μ_{i1} was composed in the subsequent form:

$$\mu_{i1} = \begin{cases} 1, & z_i < c_{i1} \\ \frac{d_{i1} - z_i}{d_{i1} - c_{i1}}, & c_{i1} \leq z_i \leq d_{i1} \\ 0, & z_i > d_{i1}. \end{cases} \quad (6)$$

In these equations, c_{i1} and d_{i1} stand for parameters which are determined based on experimental data combined with expert’s knowledge. Additionally, for the output variable of the FIS three membership functions are defined [21]. In the given situation, this could be used directly to form a compensation factor (Fig. 7),

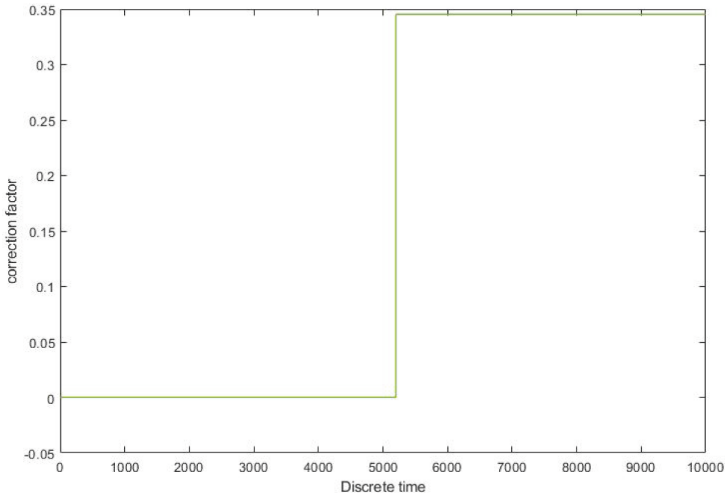


Fig. 7. Correction Factor.

The integration between FTD and FTC is already apparent on the level of requirements (compare Fig. 4). The exploration of possible and probable faults is a main prerequisite for the design of the FVA and connects FTD and FTC. On the functional level, the principle of the FVA can be shown in a relation-oriented function model (Fig. 8).

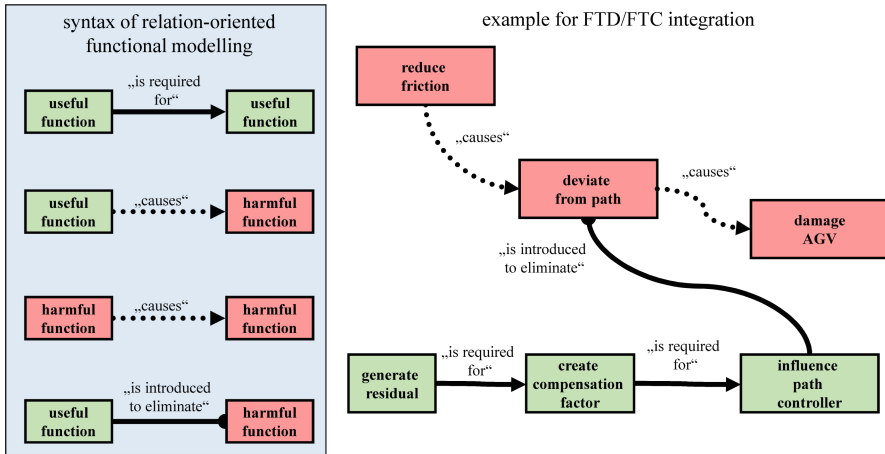


Fig. 8. Relation-oriented function model of the FVA.

On the left side of Fig. 8 a syntax explanation is given; this syntax is mainly based on [8]. On the right side, an example for a function model of the FTD/FTC integration describing the given functionality of the FVA is given.

5 Summary and Outlook

This paper explained important aspects of the integration of FTD and FTC. This integration is very important for efficient and effective fault accommodation in complex technical systems. The combination of FTD and FTC fosters the synthesis of technical systems which are more robust concerning the behaviour in the presence of faults and can actively react to faults in a coordinated and sensible manner. Through the integration of FTD and FTC, the reliability, safety and overall efficiency of technical systems can be greatly enhanced. The underlying examples were the design and control of AGVs. As a prominent example a fuzzy virtual actuator was explained. Several interesting fields for future research could be identified. One major field is the co-simulation of kinematic and control processes. Another major field are integrated evaluation and optimization methods.

Acknowledgment. A part of the work was supported by the National Science Centre of Poland under Grant: UMO-2017/27/B/ST7/00620. A part of the research work was carried out in the scope of the project “Automatisierter Entwurf eines geometrischen und kinetischen digitalen Zwillings einer Rohbaufertigungsanlage für die Virtuelle Inbetriebnahme (TWIN)”, which is funded by the German Federal Ministry of Education and Research.

References

1. Alyoussef, F., Akrad, A., Sehab, R., Morel, C., Kaya, I.: Velocity sensor fault-tolerant controller for induction machine using intelligent voting algorithm. *Energies* **15**(9), 3084 (2022)
2. Berx, N., Decré, W., Morag, I., Chemweno, P., Pintelon, L.: Identification and classification of risk factors for human-robot collaboration from a system-wide perspective. *Comput. Ind. Eng.* **163**, 107827 (2022)
3. Blanke, M., Kinnaert, M., Lunze, J., Staroswiecki, M.: *Diagnosis and Fault-Tolerant Control*. Springer-Verlag, New York (2016)
4. Dubrova, E.: *Fault-Tolerant Design*. Springer-Verlag, New York (2013)
5. Ehrlenspiel, K., Meerkamm, H.: *Integrierte Produktentwicklung: Denkabläufe, Methodeneinsatz, Zusammenarbeit*, 6., vollständig überarbeitete und erweiterte Auflage. München Wien, (2017)
6. Elwert, M., Ramsaier, M., Eisenbart, B., Stetter, R., Till, M., Rudolph, S.: Digital function modeling in graph-based design languages. *Appl. Sci.* **12**(11) (2022)
7. Gillijns, S., De Moor, B.: Unbiased minimum-variance input and state estimation for linear discrete-time systems. *Automatica* **43**, 111–116 (2007)
8. Herb, R., Herb, T.: and Veit Kohnhauser. *Triz. Der systematische Weg zur Innovation*, Landsberg (2000)

9. Hua, L., Zhang, J., Li, D., Xi, X., Asif Shah, M.: Sensor fault diagnosis and fault tolerant control of quadrotor UAV based on genetic algorithm. *J. Sens.* (2022)
10. Khoury, B., Nejjari, F., Puig, V.: Reliability-aware zonotopic tube-based model predictive control of a drinking water network. *Int. J. Appl. Math. Comput. Sci.* **32**(2), 197–211 (2022)
11. Kościelny, J.M., Bartyś, M., Syfert, M., Szyber, A.: A graph theory-based approach to the description of the process and the diagnostic system. *Int. J. Appl. Math. Comput. Sci.*, **32**(2), 213–227 (2022)
12. Majdzik, P., Witczak, M., Lipiec, B., Banaszak, Z.: Integrated fault-tolerant control of assembly and automated guided vehicle-based transportation layers. *Int. J. Comput. Integr. Manuf.* **35**(4–5), 409–426 (2022)
13. Mrugalska, B., Stetter, R.: Health-aware model-predictive control of a cooperative agv-based production system. *Sensors* **19**(3), (2019)
14. Pahl, G., Beitz, W., Feldhusen, J., Grote, K.H.: *Engineering Design: a systematic Approach*. Springer-Verlag, (2007)
15. Rotondo, D., Ponsart, J.-C., Fatiha Nejjari, Theilliol, D., Puig, V.: Virtual actuator-based FTC for LPV systems with saturating actuators and FDI delays. In: 2016 3rd Conference on Control and Fault-Tolerant Systems (SysTol), pp. 831–837. IEEE (2016)
16. Rotondo, D., Ponsart, J.-C., Theilliol, D., Nejjari, F., Puig, V.: A virtual actuator approach for the fault tolerant control of unstable linear systems subject to actuator saturation and fault isolation delay. *Annual Rev. Control* **39**, 68–80 (2015)
17. Rotondo, D., Puig, V., Nejjari, F.: A virtual actuator approach for fault tolerant control of switching LPV systems. *IFAC Proc. Vol.* **47**(3), 11667–11672 (2014)
18. Seron, M., De Doná, j., Richter, J.H.: Bank of virtual actuators for fault tolerant control. *IFAC Proc. Vol.* **44**(1), 5436–5441 (2011)
19. Srinivasarengan, K., Ragot, J., Aubrun, C., Maquin, D.: Parameter identifiability for nonlinear LPV models. *Int. J. Appl. Math. Comput. Sci.* **32**(2), 255–269 (2022)
20. Stetter, R.: *Fault-Tolerant Design and Control of Automated Vehicles and Processes*. SSDC, vol. 201. Springer, Cham (2020). <https://doi.org/10.1007/978-3-030-12846-3>
21. Stetter, R.: A virtual fuzzy actuator for the fault-tolerant control of a rescue vehicle. In: *Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, Glasgow, UK (2020)
22. Stetter, R.: Algorithms and methods for the fault-tolerant design of an automated guided vehicle. *Sensors* **22**(12) (2022)
23. Stetter, R., Göser, R., Gresser, S., Till, M., Witczak, M.: Fault-tolerant design for increasing the reliability of an autonomous driving gear shifting system. *Eksplotacja i Niezawodność*, **22**(3) (2020)
24. Stetter, R., Witczak, M., Pazera, N.: Virtual diagnostic sensors design for an automated guided vehicle. *Appl. Sci.* **8**(5) (2018)
25. Wang, R.C., Edgar, T.F., Baldea, M., Nixon, M., Wojsznis, W., Dunia, R.: Process fault detection using time-explicit kivi diagrams. *AIChE J.* **61**(12), 4277–4293 (2015)
26. Witczak, M.: Fault diagnosis and fault-tolerant control strategies for non-linear systems. In: *Lecture Notes in Electrical Engineering*, Vol. 266. Springer International Publishing, Heidelberg, Germany (2014). <https://doi.org/10.1007/978-3-319-03014-2>
27. Zhang, K., Li, H., Cao, S., Yang, C., Sun, F., Wang, Z.: Motor current signal analysis using hypergraph neural networks for fault diagnosis of electromechanical system. *Measurement* **201**, 111697 (2022)



Problems of Using Eddy Current Arrays NDT

Iuliia Lysenko¹ (✉), Yuriy Kuts¹, Valentyn Uchanin², Yordan Mirchev³,
and Alexander Alexiev³

¹ National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”, Kyiv,
Ukraine

j.lysenko@kpi.ua

² Karpenko Physico-Mechanical Institute of the National Academy of Sciences of Ukraine,
Lviv, Ukraine

³ Institute of Mechanics at the Bulgarian Academy of Sciences, Sofia, Bulgaria

Abstract. Eddy current arrays (ECAs) have become increasingly popular in non-destructive testing (NDT) due to their ability to provide fast and accurate inspection results. However, like any other NDT method, ECAs have their limitations and drawbacks. This article discusses some of the problems associated with using ECAs for NDT, including the effect of array size and geometry on inspection performance, the influence of material properties on signal interpretation, and the challenge of optimizing inspection parameters for different inspection scenarios. The article also highlights recent advances in ECA technology that aim to overcome some of these problems, including the use of advanced signal processing techniques and the development of novel array designs. Overall, this article provides a comprehensive overview of the challenges associated with using ECAs for NDT and offers insights into the future directions of ECA research and development.

Keywords: Eddy Current Technology · Non-Destructive Testing · Inspection Process · Eddy Current Arrays · Material Properties · Signal Processing

1 Introduction

Eddy current arrays (ECAs) are widely used in non-destructive testing (NDT) due to their ability to provide fast and accurate inspection results. However, several challenges associated with using ECAs for NDT need to be addressed to improve inspection performance.

One of the main problems with ECAs is the effect of array size and geometry on inspection performance. The size and shape of the ECA affect the sensitivity and resolution of the inspection, and optimizing these parameters for a specific inspection scenario is a challenge. Additionally, the use of large ECAs can result in increased noise levels and decreased inspection speed, which impacts the overall efficiency of the inspection process.

Another problem with ECAs is the influence of material properties on signal interpretation [1, 2]. The material composition, structure, and thickness affect to eddy current

response, making it difficult to interpret the inspection results accurately. Additionally, the presence of defects, such as cracks or corrosion, complicates the signal interpretation, makes it challenging to distinguish between relevant and irrelevant signals.

Finally, optimizing the inspection parameters for different inspection scenarios is a hard process. Different inspection scenarios require different inspection frequencies, coil configurations, excitation modes of probes, and signal processing techniques [3, 4]. Additionally, complex geometry or surface irregularities lead to further complicate the inspection process, requiring the development of new inspection techniques and strategies [5].

To address these challenges, researchers are working on developing new signal-processing techniques and novel ECA designs. For example, they develop advanced signal processing algorithms to help improve the signal-to-noise ratio, reduce false positives, and enhance defect detection capabilities [6–8]. Additionally, the development of new ECA designs, such as flexible and conformable arrays, allows improving inspection performance on complex geometries and irregular surfaces [9].

Overall, while ECAs are a promising NDT technique, several challenges associated with their use need to be improved in their inspection performance. Current research and development in this area will likely lead to improved ECA technology and enhanced inspection capabilities.

This paper examines the importance of understanding the processes of signal formation and processing as an integral part of the developed tuning methods and recommendations for working with the equipment.

2 Research and Discuss the Influence Factors on NDT Performance

2.1 Effect of Array Size and Geometry

Practice shows the sensitivity and resolution of the inspection are affected by the size and shape of the array, making it challenging to optimize these parameters for a specific inspection scenario. Additionally, using large arrays results in increased noise levels and decreased inspection speed, which negatively impact the efficiency of the inspection process. Therefore, optimizing the array size and geometry is crucial for improving the accuracy and efficiency of non-destructive inspections.

Research on the influence of geometry and size of arrays on sensitivity and resolution of inspection focused on providing optimal design of magnetic arrays for defect detection. Researchers proposed to use of computer modeling to analyze different geometries and sizes of magnetic arrays to determine which design is most suitable for different types of defects [4, 10]. The research shows the optimization analysis of the parameter settings of defect detection through simulation modeling is needed to design the probe design and excitation setting in the actual detection process [10].

The simulation results reduce the cost of the probe designer's research process and lead to the production of the finished product. The article [11] presents a study of a flexible planar matrix of eddy current probes for monitoring the presence of microcracks in critical parts of aircraft. The production of such a probe became possible thanks to missed preliminary simulations and a permanent feature of the probe - both exciting

and sensitive coils are etched on polyimide films using flexible printed circuit board technology. Experimental results showed that the probe matrix is sensitive to microcracks and is able to determine the crack length with good accuracy [11, 12].

However, the main influence on the signal of array probes represented by coils is from their mutual influence on each other. Physical processes that occur when two coils are placed quite close can be described by the laws of magnetic field propagation and illustrated by the equivalent scheme in Fig. 1 (where u_G – generator, R – the generator output resistance, R_1 – the coil active resistance, R_1 and L_1 – the excitation coil active resistance and inductance, R_2 and L_2 – the receive coil active resistance and inductance, R_3 and L_3 – the testing sample active resistance and inductance, M_{12} , M_{23} , M_{13} – the magnetic coupling coefficient, i_1 , i_2 and i_3 – the currents in the branches) [13, 14].

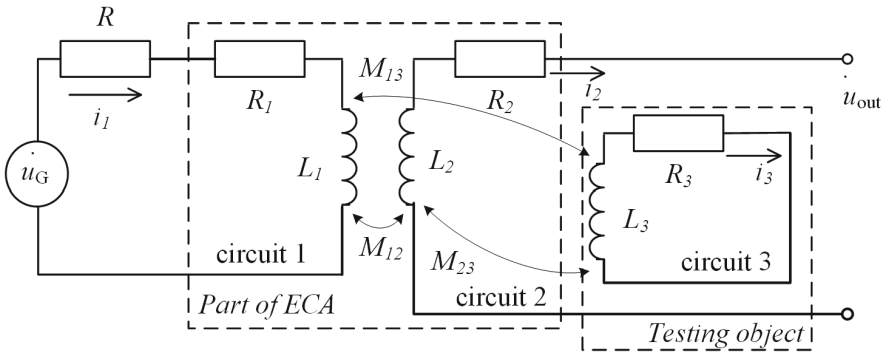


Fig. 1. Equivalent circuit of the system of “Excitation coil - receiving coil - testing object”

Figure 1 Shows one pair of coils that are working together in an array probe at one time to excite and receive signals during the testing. Obviously, the analysis of array probes for ECT is a mathematically complex process, but it is necessary. After all, for the optimal choice of the dimensions and geometry of the probe, it is important to determine the possible mutual absorption of the coils and to predict its consideration during signal processing.

2.2 Effect of Inspection Parameters Choice

The challenge of optimizing inspection parameters for different inspection scenarios in eddy current arrays NDT lies in the need to balance the trade-off between sensitivity and resolution. Optimizing the inspection parameters such as the frequency, excitation voltage, and probe geometry for a specific inspection scenario affect the performance of eddy current arrays. However, there is no one-size-fits-all solution, as each inspection scenario presents different challenges that require a tailored approach. However, the specialist of NDT has to clearly understand this and, if necessary, be able to perform equipment settings not included in the procedures to achieve the highest quality result.

As part of the research, an ECA probe was used to diagnose a two-layer sample for the presence of non-adhesion between the layers, and the technological instruction for using the ECA probe was validated.

For example, Fig. 2 and 3 show the testing results from one sample with the same inspection mode, but different signal processing settings were used, and, as a result, different quality results were obtained.

The excitation frequency was selected at 19 kHz in both cases. In the first variant, the gain settings were 85 dB (vertical 100 dB), while in the second it was 80 dB (vertical 100 dB). The difference in scanning angle with respect to the edge of the object in the first and second cases was 5°. The amplitude and phase of the signal voltage were: $A_{max} = 10.3 \text{ V}$ and $\varphi = -4.7^\circ$ for the case in Fig. 2 and $A_{max} = 7.6 \text{ V}$ and $\varphi = 9.7^\circ$ for Fig. 3.

Scanning was performed using an eddy current array (Fig. 4), which contained 64 transducers in two rows (32 each). The size of the probe was $64 \times 32 \text{ mm}$. Such an array of transducers allows for the inspection of a larger area of the material in a single pass and is capable of stable operation over a wide temperature range. The design of the probe allows for its use in both manual and automatic testing.

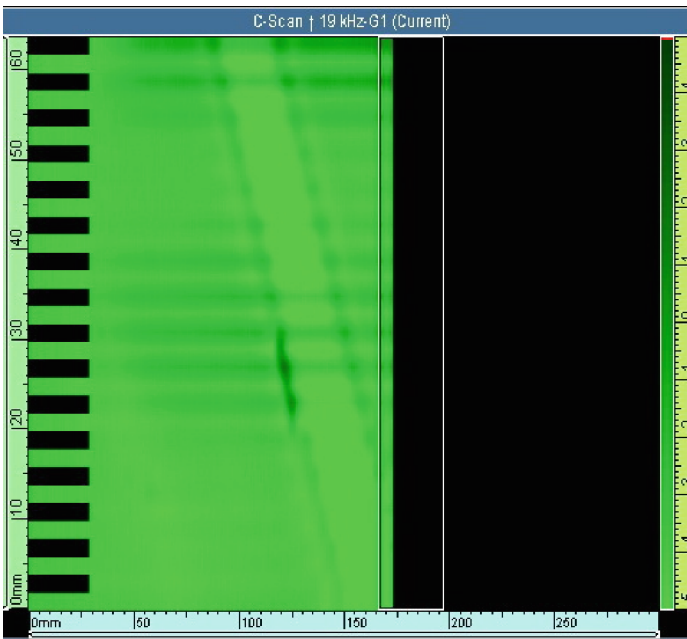


Fig. 2. Testing results with recommended display settings

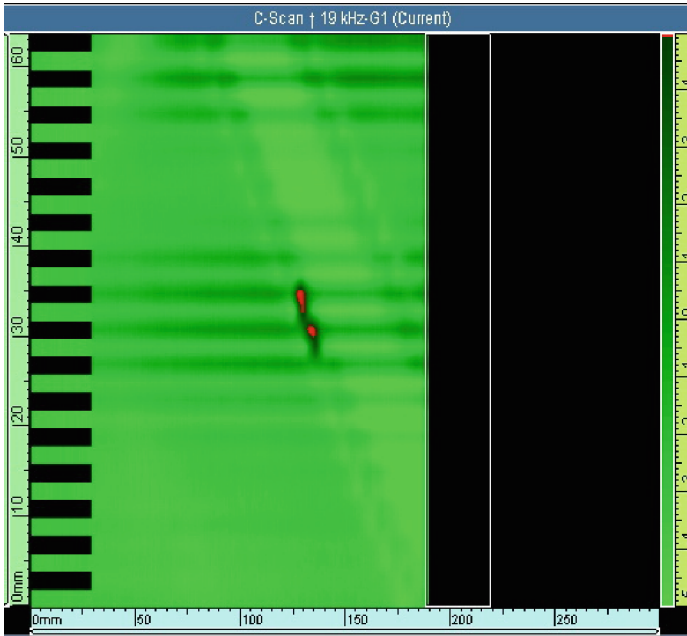


Fig. 3. Testing results with specific display settings

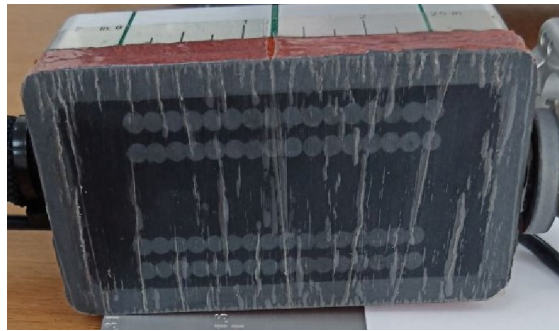


Fig. 4. Eddy current array probe

In those cases, the aluminum two layers plate, which has some not glued area between the layers, was scanned by ECA, and after it, the different display settings were used. It is hard to identify the defect with sufficiently high accuracy using Fig. 2. Also, the size and shape of a defect are not obvious on this scan. Figure 3 gives more good identification even though the processing was carried out with a slight deviation from the procedure.

Therefore, finding the optimal inspection parameters and signal processing (presenting) method for each specific inspection scenario requires a thorough understanding of the material properties, the type of defect being targeted, and the available eddy current array equipment. It also requires expertise in selecting and adjusting the appropriate

parameters to maximize the performance of the system while minimizing false positives and false negatives.

3 Conclusions

The article discusses the challenges associated with using eddy current arrays (ECAs) for non-destructive testing (NDT) and provides insights into future directions of ECA research and development. There are identified several problems with ECAs, including the effect of array size and geometry on inspection performance, the influence of material properties on signal interpretation, and the challenge of optimizing inspection parameters for different inspection scenarios. To overcome these challenges, it is proposed the development of new signal processing techniques and novel ECA designs. The article also highlights recent research that uses computer modeling to analyze different geometries and sizes of magnetic arrays to optimize their design for defect detection. The results of testing the object with some not glued areas inside are presented and it demonstrated the influence of specific display settings on the visual perception of the results. Overall, the article provides a comprehensive overview of the challenges associated with using ECAs for NDT and offers insights into the future directions of ECA research and development.

This article is practically oriented and describes the author's experience in using ECAs. The experiments in this work were conducted using harmonic excitation of the probe, but in the future, the use of an ECA probe for pulse excitation is planned. As a result of the research, it was found that the testing result depends on the scanning angle (scanning direction with respect to the defect such as non-adhesion type) and the settings of the display system. The studies were conducted according to the manufacturer's specifications, which did not take into account the peculiarities of using the probes in practice. In the future, it is necessary to develop a testing methodology that would take into account the identified shortcomings.

Acknowledgements. This work has been accomplished with financial support by the Grant № BG05M20P001–1.002–0011 „Establishment and development of a Center for Competence in Mechatronics and Clean Technologies MIRACle (mechatronics, innovation, robotics, automation, clean technologies)”, financed by the Science and Education for Smart Growth Operational Program (2014–2020) and co-financed by the European Union through the European Structural and Investment Funds.

References

1. Hocking, J.G.: Eddy current inspection of complex components: the problem of signal interference. *Insight-Non-Destructive Test. Cond. Monit.* **54**(4), 203–206 (2012)
2. Libby, H.L.: *Introduction to Electromagnetic Nondestructive Test Methods*. Wiley-Interscience New York (1971)
3. Aviña-Cervantes, J.G., Rodríguez-Castellanos, A.: Non-destructive testing by eddy current: new trends and future directions. *J. Market. Res.* **7**(4), 389–395 (2018)

4. Thibault, S., Maldague, X.: Advanced eddy current array technology for surface crack detection. In: Proceedings of the 8th International Conference on Barkhausen Noise and Micromagnetic Testing (ICBNMT), pp. 1–8 (2013)
5. Lysenko, I., Uchanin, V., Petryk, V., Kuts, Y., Protasov, A., Alexiev, A.: Intelligent automated eddy current system for monitoring the aircraft structure condition. In: 2022 IEEE 3rd International Conference on System Analysis & Intelligent Computing (SAIC), Kyiv, Ukraine, pp. 1–5 (2022)
6. Nesbitt, A., Drinkwater, B.W.: Detection of small defects using eddy current arrays. *NDT E Int.* **86**, 26–34 (2017)
7. Niu, L., Tian, G.Y.: A novel approach to eddy current array probe optimization for non-destructive testing. *Meas. Sci. Technol.* **26**(4), 045009 (2015)
8. Uchanin, V.M., Ivashchenko, K.A.: Detection of defects of structures from ferromagnetic steel through the layer of anticorrosion cover without removal [in Ukrainian]. *Methods Dev. Qual. Control* **1**(46), 5–14 (2021)
9. Liu, X., Wang, L., Li, Y., Li, S., Li, C.: Application of eddy current testing in material evaluation and defect detection. *J. Mater. Sci. Technol.* **50**, 114–126 (2020)
10. Deng, W., Bao, J., Luo, S., Xiong, X.: Simulation analysis of eddy current testing parameters for surface and subsurface defect detection of aviation aluminum alloy plate. *J. Sens.* **2022**(8111998), 1–11 (2022)
11. Xie, R., et al.: Fatigue crack length sizing using a novel flexible eddy current sensor Array. *Sensors* **15**, 32138–32151 (2015)
12. Machado, M.A., Rosado, L.S., Santos, T.G.: Shaping eddy currents for non-destructive testing using additive manufactured magnetic substrates. *J. Nondestr. Eval.* **41**(3), 1–24 (2022)
13. Alexander, C., Sadiku, M.: *Alexander Fundamentals of Electric Circuits*, 4th edn. The McGraw-Hill Companies Inc, NY (2009)
14. Zhong, M., Kuts, Y., Kochan, O., Lysenko, I., Levchenko, O., Vlach-Vyhrynovska, H.: Using signal phase in computerized systems of non-destructive testing. *Measur. Sci. Rev.* **22**(1), 32–43 (2022)



Calibration of a High Sampling Frequency MEMS-Based Vibration Measurement System

Muhammad Ahsan^(✉) and Dariusz Bismor

Department of Measurements and Control Systems, Silesian University of Technology, 44-100 Gliwice, Poland
{Muhammad.Ahsan,Dariusz.Bismor}@polsl.pl, ahsanmuhammad@aol.com

Abstract. Vibration characteristics of a low-cost measurement system based on the ADXL100x micro-electromechanical system (MEMS) sensors are presented in this paper. The discussed measurement system consists of two ADXL100x MEMS accelerometers, interfaced with a BeagleBone Black microcontroller. The accelerometers were mounted on a vibration exciter equipped with a reference accelerometer. The vibration exciter was feeded using a function generator with various frequencies and acceleration signals. Vibration characteristics were investigated by comparing reference and recorded frequencies, sensitivity illustration over various frequencies, and sensitivity effects on different accelerations. The results demonstrate that the ADXL100x MEMS sensor is efficient and suitable for high-sensitivity applications.

1 Introduction

Recent technological advancements have facilitated the extensive manufacturing of micro-electromechanical system (MEMS) sensors for a wide range of applications. MEMS technology is not only shrinking in size, weight, and power, but it is also improving in functionality. Initially, MEMS sensors were utilized exclusively in toys, mobile phones, laptops, or automotive components, but they are now appropriate for gravity measurements, navigation systems, and other extremely sensitive applications because of improved measurement quality and technological progress [1–3].

Many researchers have reported on the performance of MEMS-based systems in the literature. In [4, 5], piezoelectric and MEMS accelerometers were utilized to capture vibration signals from combustion engines, and vibration-based comparison and analysis were performed. The low-cost MEMS accelerometer ADZL001-70 was interfaced with the STM32 microcontroller to develop a vibration measuring prototype in [6], however, the recorded signals were exposed to significant noise due to the low-quality vibration sensor. Furthermore, no research on sensing performance was provided. Calibrations of two accelerometers, piezoelectric PCB 338B35 and capacitive ADXL 202, were done and the results were compared in [7]. An experimental investigation employing MEMS accelerometers for

condition monitoring of a CNC machine in a typical industrial environmental workshop was completed in [8]. In [9], a vibrating beam MEMS accelerometer was developed for gravity and seismic measurements. The vibration signals are exposed to high noise when acquired from real-world industrial machines and therefore advanced signal processing techniques are applied to enhance the signal-to-noise ratio. In [10, 11], a dynamical bandpass filter was constructed and the signal-to-noise ratio was improved to diagnose the fault frequencies from the faulty vibration signals.

Health monitoring through vibration signals is critical for detecting faults in industrial machinery [10]. Traditional piezoelectric accelerometers are excellent vibration-measuring devices, but their cost is significant, especially when numerous sensors are required to collect vibration data from different places. MEMS accelerometers, on the other hand, are quite inexpensive, costing just 5 to 10% of the price of traditional piezoelectric accelerometers [4]. However, manufacturing problems such as excessive internal noise, generated asymmetric structures, misalignment of actuation mechanisms, and deviations of the center of mass from the geometric center affect the performance of MEMS accelerometers. As a consequence, it is critical to construct an efficient solution that makes use of an appropriate and cost-effective MEMS-based vibration measuring system.

Motivated by the above literature review, this research presents a vibration monitoring system based on MEMS accelerometers. MEMS accelerometers are a low-cost alternative to traditional piezoelectric accelerometers. Two ADXL100x accelerometers are interfaced with the BeagleBone Black controller board in the proposed system, and vibration signals at various frequencies and accelerations are recorded. To drive the connected vibration exciter, a function generator is used to create the various vibration signals, and both accelerometers are installed onto the vibration exciter to record the vibration signals. The primary contribution of this work is to examine the characteristics of both accelerometers. Three types of vibration characteristics are illustrated: reference frequencies vs recorded frequencies, sensitivity at various frequencies, and sensitivity at various accelerations. Characteristics of both accelerometers show efficient results, and it is concluded that ADXL100x accelerometers are the best and most cost-effective alternative solution to any conventional solution.

The rest of the paper is organized as follows: Sect. 2 comprises on the experimental setup and data recording. This section illustrates the hardware used to perform the experiment and contains the recorded vibration signals at different frequencies and acceleration values. Section 3 includes the results concluded for vibration characteristics of ADXL100x MEMS sensors. Finally, Sect. 4 includes the conclusion of this research work.

2 Experimental Setup and Data Recording

The block diagram of the experimental setup is illustrated in Fig. 1. The setup consists of a function generator, power amplifier, vibration exciter, measuring amplifier, MEMS accelerometers, and BeagleBone Black. The description of each piece of hardware used in this experiment is given below in the Table 1:

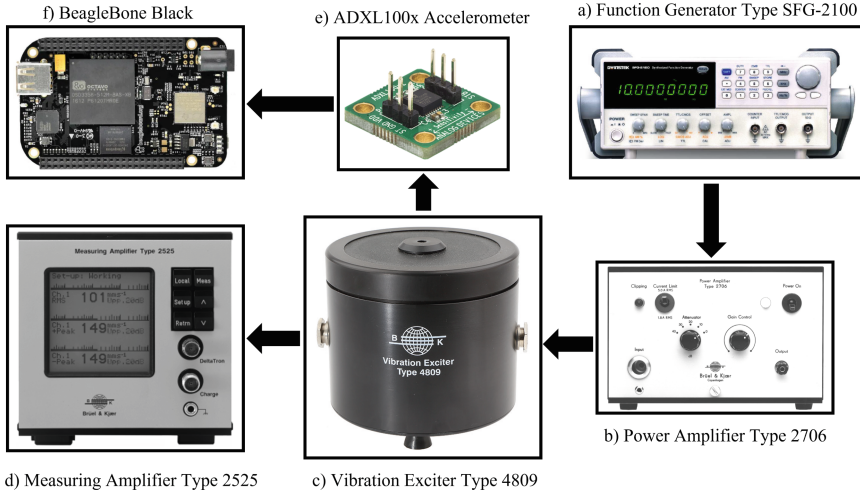


Fig. 1. Block diagram of the experimental setup.

In the experimental setup, sinusoidal signals with different frequencies were generated by the function generator and the attached vibration exciter produce vertical vibrations. Two ADXL100x accelerometers interfaced with the BeagleBone Black were mounted on the vibration exciter to measure and record the vibration signals. The recorded vibration signals were stored on the SSD card in BeagleBone Black and to study the vibration characteristic, signal processing methods were applied to the recorded data in MATLAB.

The experiment was repeated multiple times with different frequencies and accelerations. Figure 2a shows the graph of the first dataset recorded at multiple reference frequencies and constant acceleration of 1 m/s^2 . Whereas Fig. 2b shows

Table 1. Hardware type and description

Hardware	Type	Purpose
Function Generator	SFG-2100	Generate vibration signals with different frequencies and amplitudes
Power Amplifier	Type-2706	Enhance the power of the generated vibration signals by function generator
Vibration Exciter	Type-4809	Produce vibrations at specific frequency and amplitude given by the input vibration signal
Measuring Amplifier	Type-2525	Display measurement data as well as selecting setup and measurement parameters
MEMS Accelerometer	ADXL100x	Sense the vibration signals from the vibration exciter and send an output to the microcontroller
BeagleBone Black		Record the input vibration signal and save it to the SSD card for further processing

the graph of the second dataset recorded at 500 Hz, 1 kHz, 2 kHz, and 5 kHz with different accelerations.

Figure 3 shows a portion of recorded sinusoidal signals with different reference frequencies. From the figure, it can be visualized that the frequencies of the time-domain signals are different and the exact frequencies are depicted in the corresponding frequency-domain representation in Fig. 3. Furthermore, it can also be seen that low-frequency signals are more distorted than high-frequency signals.

3 Results

The acquired datasets described in the previous section were used to illustrate the vibration characteristics of the low-cost ADXL100x accelerometers. Two accelerometers were mounted on the vibration exciter simultaneously to record the vibration datasets. Two datasets were constructed: the first dataset was recorded at different frequencies with constant acceleration as shown in Fig. 2a whereas the second dataset consists of a vector of four frequencies i.e., 0.5 kHz, 1 kHz, 2 kHz, and 5 kHz and each frequency data was recorded with different level of acceleration values as shown in Fig. 2b. Three types of characteristics were computed from the recorded datasets and listed as follows:

- Characteristic # 1: Frequency characteristic includes the relationship between reference frequencies and recorded frequencies.
- Characteristic # 2: Sensitivity characteristic at different frequencies consists of a relationship between sensitivity and different frequencies.
- Characteristic # 3: Sensitivity characteristic at different accelerations consists of a relationship between sensitivity and different accelerations.

To compute the reference frequencies, a signal processing method called fast frequency transform (FFT) was implemented on the time-domain recorded signals. Figure 4 depicts the first vibration characteristic that shows the relationship between reference and acquired frequencies. The linear graph between reference frequencies and recorded frequencies shows the efficiency of the low-cost ADXL100x accelerometer. Figure 5 illustrates the visualization of the error between the reference frequency and measured frequency signals. In Fig. 5, a dotted line representing the upper bound limit and the lower bound limit is plotted at 1.5 and 1.7, which means the error signal lies within a numerical value of 0.2. This shows the performance of the low-cost ADXL100x accelerometer.

Furthermore, to examine the second and third characteristics, sensitivity was computed using the following equation:

$$S_i = \frac{\text{RMS}(x(f_i))}{g_i} \quad (1)$$

where $x(f_i)$ is the recorded time-domain vibration signal at a specific frequency f_i with $i = 1, 2, 3, \dots, n$, g_i is the acceleration of the vibration signals, $\text{RMS}(\cdot)$

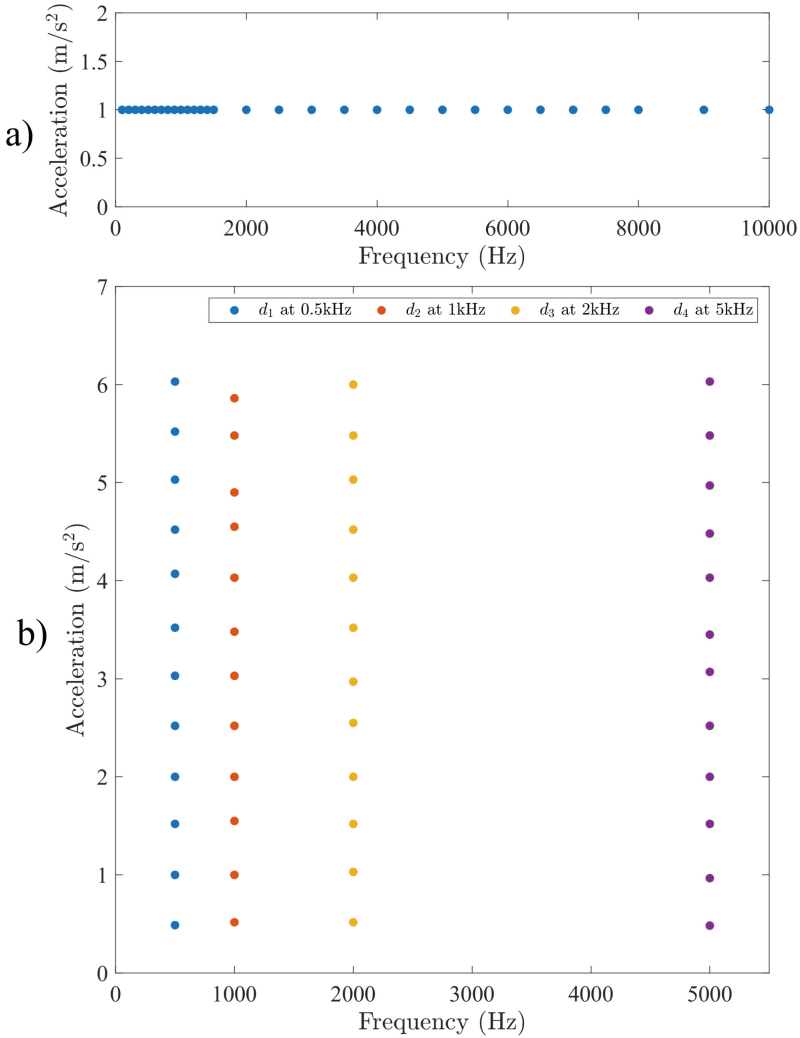


Fig. 2. Both datasets: a) first dataset at multiple reference frequencies and constant acceleration, and b) second dataset recorded at 500 Hz, 1 kHz, 2 kHz, and 5 kHz reference frequencies and different accelerations.

is the root-mean-square of the signal, and S_i is the sensitivity index. To plot the sensitivity at different frequencies, the first dataset was used. Figure 6 shows the sensitivities of both accelerometers at different frequencies. The upper and lower dotted lines show the limits for the sensitivity index. Furthermore, it can be seen that the sensitivities of both accelerometers are almost similar.

The second dataset was recorded keeping the constant frequency and varying the acceleration values. The sensitivity for each signal was computed using the

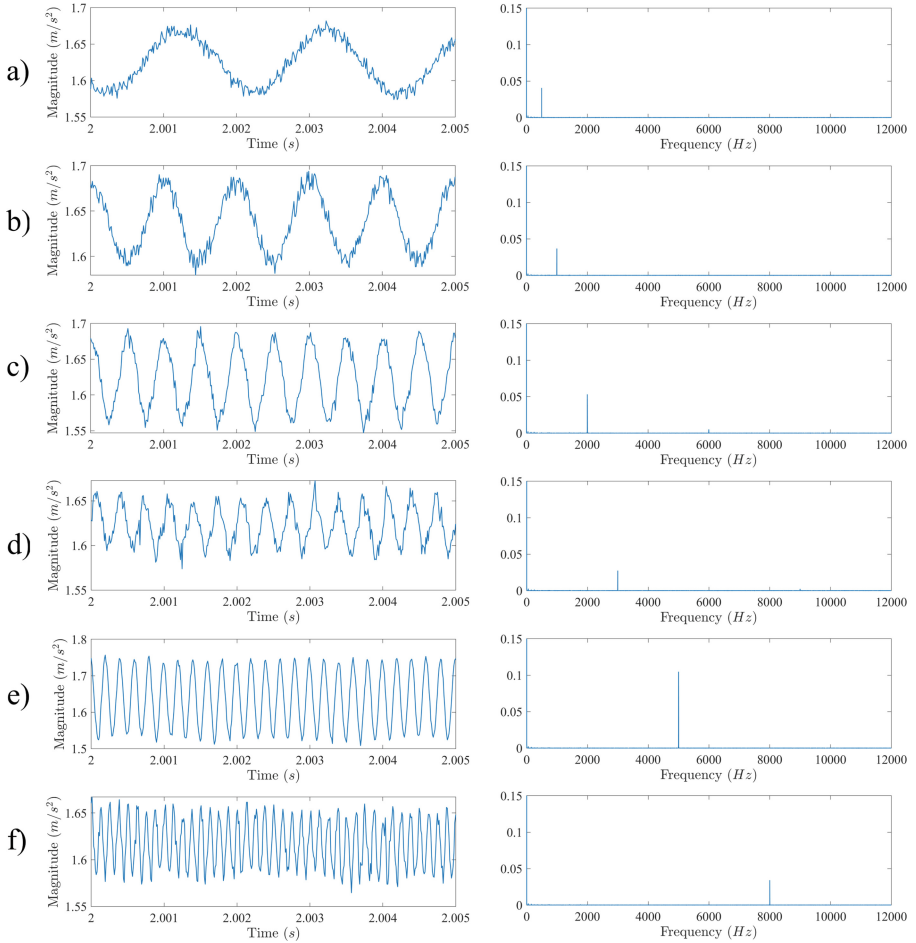


Fig. 3. Visualization of time-domain and corresponding frequency-domain signals recorded at multiple frequencies: a) 500 Hz, b) 1 kHz, c) 2 kHz, d) 3 kHz, e) 5 kHz, and f) 8 kHz.

equation (1). To illustrate the third characteristic a plot between sensitivity and acceleration was constructed for both accelerometers as shown in Fig. 7. The graphs show that the sensitivity is inversely proportional to the acceleration. All the results of vibration characteristics show that the low-cost ADXL100x accelerometers are efficient for sensitive applications and diagnose the frequencies accurately.

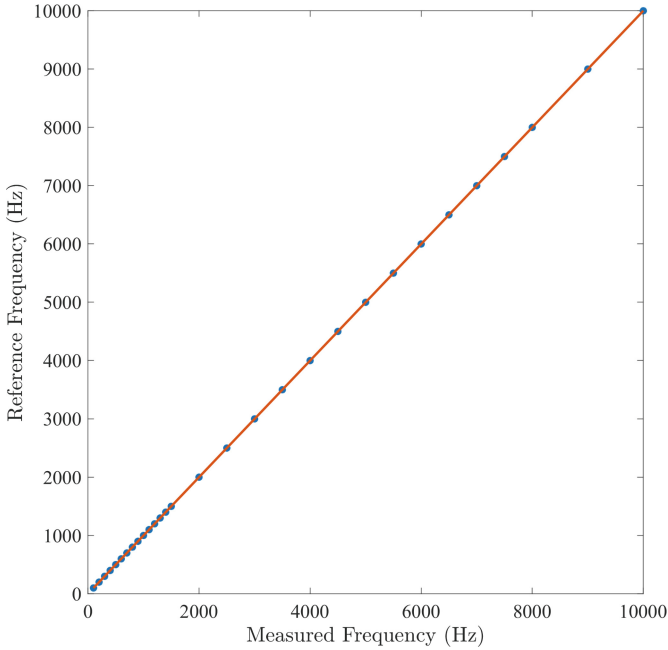


Fig. 4. Relationship graph between reference frequencies of the vibration shaker and measured frequencies of the ADXL100x accelerometer.

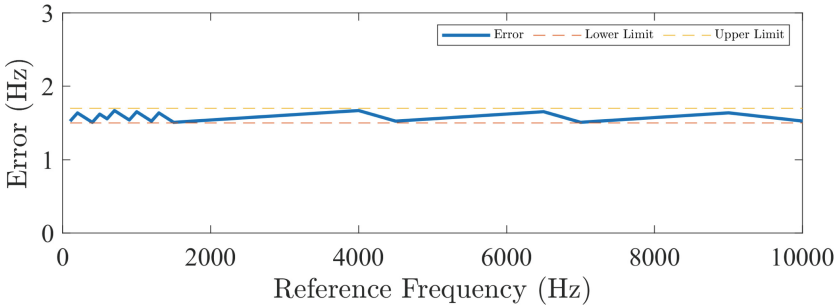


Fig. 5. Visualisation of errors between reference frequencies of the vibration shaker and measured frequencies of the ADXL100x accelerometer.

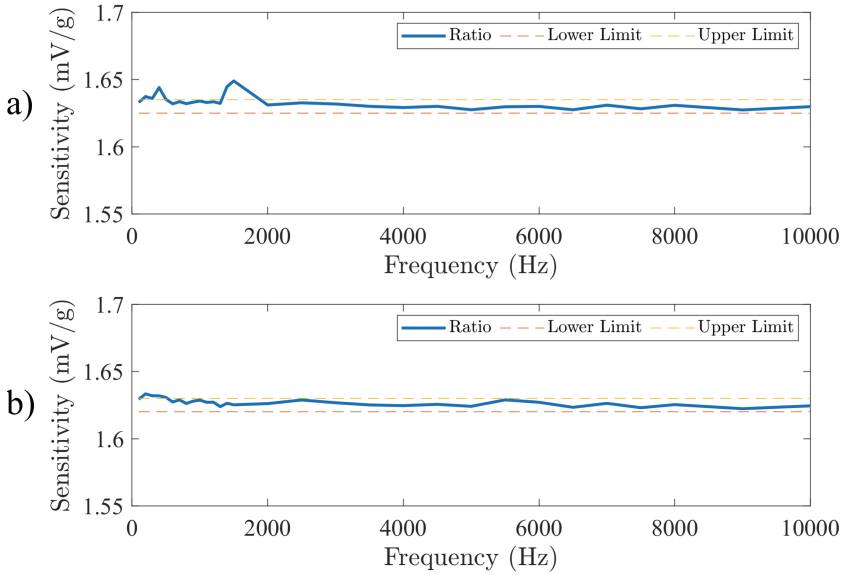


Fig. 6. Sensitivities of both accelerometers at different frequencies a) sensitivity of the first accelerometer and b) sensitivity of the second accelerometer.

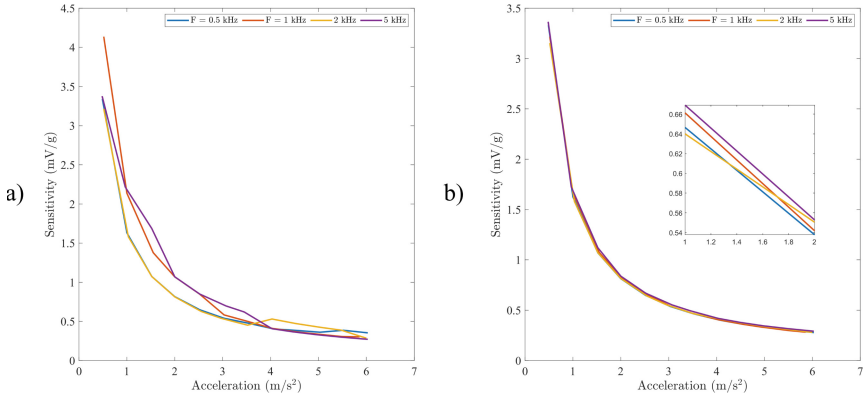


Fig. 7. Characteristic of sensitivity versus acceleration at different frequencies: a) characteristic of the first accelerometer, and b) characteristic of the second accelerometer.

4 Conclusion

In conclusion, the paper presents a comprehensive study of a MEMS-based vibration monitoring system that utilizes the ADXL100x accelerometer and a BeagleBone Black controller board. The results show that the system is efficient in recording vibration signals at various frequencies and accelerations. The vibra-

tion characteristics, such as reference frequencies vs recorded frequencies, sensitivity at various frequencies, and sensitivity with various accelerations, demonstrate the efficiency of the MEMS-based system.

Overall, the study suggests that MEMS-based vibration monitoring systems are a promising and cost-effective alternative to conventional solutions. With their small size, low cost, and high performance, MEMS accelerometers have the potential to revolutionize the field of vibration monitoring. Therefore, it is recommended to further explore and develop MEMS-based systems for vibration monitoring applications.

References

1. Pike, W.T., Standley, I.M., Calcutt, S.B., Mukherjee, A.G.: A broad-band silicon microseismometer with 0.25 NG/rtHZ performance. In 2018 IEEE Micro Electro Mechanical Systems (MEMS), pp. 113–116 (2018)
2. Middlemiss, R.P., Samarelli, A., Paul, D.J., Hough, J., Rowan, S., Hammond, G.D.: Measurement of the earth tides with a MEMS gravimeter. *Nature* **531**, 614–617 (2016)
3. Tang, S., et al.: A high-sensitivity mems gravimeter with a large dynamic range. *Microsyst. Nanoeng.* **5**(1), 45 (2019)
4. Bismor, D.: Analysis and comparison of vibration signals from internal combustion engine acquired using piezoelectric and mems accelerometers. *Vibr. Phys. Syst.* **30**(1 2019112), 1–8 (2019)
5. Bismor, D.: System for vehicle sound and vibration monitoring using mems sensors. In: Proceedings of the IEEE SPA 2016 Conference, pp. 50–55. Poznań, Poland (2019). Poznan University of Technology
6. Jabłoński, A., Żegleń, M., Staszewski, W., Czop, P., Barszcz, T.: How to build a vibration monitoring system on your own? In: Timofiejczuk, A., Chaari, F., Zimroz, R., Bartelmus, W., Haddar, M. (eds.) CMMNO 2016. ACM, vol. 9, pp. 111–121. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-61927-9_11
7. Buchczik, D., Pawelczyk, M.: Calibration of accelerometers using multisinusoidal excitation. In: Timofiejczuk, A., Chaari, F., Zimroz, R., Bartelmus, W., Haddar, M. (eds.) CMMNO 2016. ACM, vol. 9, pp. 213–222. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-61927-9_20
8. Albarbar, A., Mekid, A.S.S., Pietruszkiewicz, R.: Suitability of mems accelerometers for condition monitoring: an experimental study. *Sensors* **8**(2), 784–799 (2008)
9. Mustafazade, A., et al.: A vibrating beam mems accelerometer for gravity and seismic measurements. *Scientific Rep.* **10**(1), 10415 (2020)
10. Ahsan, M., Bismor, D.: Early-stage fault diagnosis for rotating element bearing using improved harmony search algorithm with different fitness functions. *IEEE Trans. Instrum. Meas.* **71**, 1–9 (2022)
11. Ahsan, M., Bismor, D.: Early-stage faults detection using harmony search algorithm and STFT-based spectral kurtosis. In: Szewczyk, R., Zieliński, C., Kaliczyńska, M. (eds.) AUTOMATION 2022. AISC, vol. 1427, pp. 75–84. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-03502-9_8

Design of Control Systems



Configurable Dynamics of Electromagnetic Suspension by Fuzzy Takagi-Sugeno Controller

Adam Krzysztof Pilat^(✉), Hubert Milanowski, Rafal Bieszczad,
and Bartiomiej Sikora

Faculty of Electrical Engineering, Automatics, Computer Science and Biomedical
Engineering, Department of Automatic Control and Robotics,
AGH - University of Science and Technology, Kraków, Poland
{ap,milan,rafbiesz,bsikora}@agh.edu.pl

Abstract. This elaboration presents the synthesis of the Takagi-Sugeno type Fuzzy Logic controller realizing the programmable parameters of the state feedback controller together with the steady state current for the active magnetic levitation system. Closed-loop dynamics was calculated precisely with respect to the requested dynamics. Two cases were considered: fixed- and variable-closed-loop dynamics. The Fuzzy Logic Controller was considered under four scenarios with linear, nonlinear, and constant in-range output. The study was supported by simulations and experimental investigations. The position of the levitating sphere stands as an illustration of the system in operation.

1 Introduction

Fuzzy systems have been used in recent years to solve control problems, where plants are poorly modeled or nonlinear. The most popular fuzzy logic controller design method is based on the user experience, when the membership functions and rules are designed manually [1]. The gain scheduling control method embedded in a form of fuzzy logic is a typical approach [1–4]. Very often, the PID controller is configured in the form of a fuzzy controller [5–7]. One can find a proposal of an optimal search for fuzzy logic controller based on the excitation and membership functions shape optimization [8]. Takagi-Sugeno type fuzzy control is one of the most popular and promising model-based control methods of this category. It is based on a ‘fuzzy partition’ of the input space and can be viewed as an expansion of the piecewise linearization method [9]. This paper describes an application of the Takagi-Sugeno (TS) controller [10] with constant output. In this method, the main idea is to define a set of linear controllers and fuzzy rules to switch between them. Two Mamdani-type fuzzy PD controllers were used in [11] to independently stabilize (without cross-coupling and gyroscopic effects) the radial position of the rotor in a bearingless induction motor. The process of adjusting both the membership functions and the rules was carried out through

experimentation. The comparison of fuzzy controller and PID controller for magnetic levitation system is presented in [7, 12]. In [13] Mamdani-based inference and defuzzification of the center of gravity were used to design the fuzzy logic controller and a metaheuristic nature-inspired algorithm was implemented for optimization of the controller parameters. In [14] Takagi-Sugeno fuzzy model is derived for the magnetic levitation ball system by the method of sector non-linearity. The integral state feedback controller was supported by model-based design. In [15] the analytical model of the magnetic levitation system with two electromagnets was linearized around several operating points, and the simulated annealing algorithm was used to obtain the Takagi-Sugeno fuzzy model of the system. In [16] a fuzzy active disturbance rejection control (ADRC) method was designed based on fuzzy control and ADRC theory. In [17] different control methods developed for the alternative control task of tracking an axial dynamic target during levitation were compared. In [18] fuzzy control system based on a group of rules was applied to an axial active magnetic bearing due to the not well-known dynamic behavior. These rules were based on experimental results of a derivative proportional controller (PD). Fuzzy logic has also found application in designing state observers. In [19, 20] sliding mode observer performance was improved by using a fuzzy controller. In [21] an adaptive fuzzy observer is proposed to improve velocity estimation in servomotor drive.

2 Motivation

The direct motivation to implement a configurable dynamics fuzzy controller is its use in stabilizing levitating objects at different levels in relation to the surface of the electromagnet. The nonlinearities of the object make it impossible to use one linear regulator for the whole operating range, because the properties of the closed system change with the changing state of the object. It is possible to use the feedback linearization method, which, however, is sensitive to errors in the identification of the real object [9]. The practical application of levitation systems, for example, in train suspensions, magnetic bearings [22, 23], requires the design of dynamic properties for specific application requirements. Moreover, the limitation of the area of movement of the levitating object determines the critical conditions in the form of the required dynamic properties in the immediate vicinity of the electromagnet, where overshoots are not allowed (see Fig. 1). Hence, aperiodic dynamic - preferably critical properties - are required. In the remainder of the levitation region, the properties of the closed system may be freely shaped, including the possibility of oscillations with the desired elastic-damping properties. Thus, fuzzy logic was selected as an effective tool for the implementation of control of nonlinear systems. The fuzzy logic control using the Takagi-Sugeno method allows for developing a nonlinear model or a nonlinear regulator based on linear models or regulators. In this work, linear regulators were designed for specific states of a linear object by setting specific dynamic properties of the closed system. To implement the plan, a methodology was developed for the automatic generation of a fuzzy controller with defined parameters of the membership function according to the desired controller parameters. Typical shapes

and distributions of membership functions were tested to determine the range of variability or constancy of the controller parameters. The magnetic levitation system steered by the designed fuzzy controller was subjected to simulation and experimental tests.

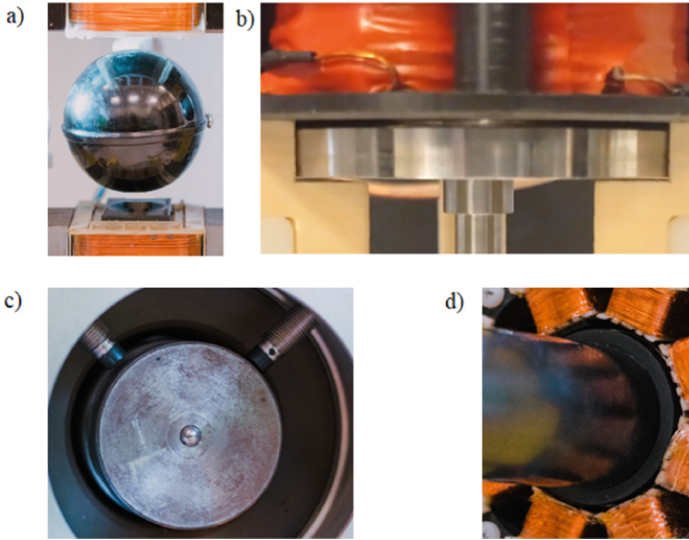


Fig. 1. Limited operation range for levitation devices (AGH, Laboratory of Magnetic Levitation, www.maglev.agh.edu.pl): a) dual active magnetic suspension, b) 6 poles axial active magnetic bearing, c) 7.5 kg rotor in radial active magnetic bearing, d) lightweight ring-type rotor in radial active magnetic bearing.

3 Theory

The idea of implementing a fuzzy controller using a set of local models consists of linearization of the nonlinear system at many operating points, selection of parameters of a linear controller for a dedicated operating point, and selection of a specific model using the rules used in the fuzzy control theory. The concept of the synthesis of the regulation system presented below is based on the theory developed by Takagi-Sugeno [10,24]. Let us consider a nonlinear system of a single axis magnetic bearing described by a k linear model based on heuristic rules:

R^i : **IF** x is F^i **then** fuzzy subsystem is described:

$$\begin{cases} \dot{x} = A_i x + B_i u_i \\ y = C_i x \end{cases} \quad (1)$$

where: R^i - rule ($i \in (1, k)$), k - number of rules, F^i - fuzzy set, (A_i, B_i, C_i) - controllable and observable local linear model, u_i - control, y - output.

The fuzzy set center is located at the linearization point and represents the range of the state variable as a membership function $\mu_i(x)$.

The normalized coefficient for the i^{th} rule is given in the following form:

$$w_i(x) = \frac{\mu_i(x)}{\sum_{i=1}^k \mu_i(x)}, \sum_{i=1}^k w_i(x) = 1 \quad (2)$$

In this case, the fuzzy dynamical model of the system can be described by (3).

$$\begin{aligned} \dot{x} &= Ax + Bu \\ y &= Cx \end{aligned} \quad (3)$$

where: A, B, C are calculated as weighted local matrices A_i, B_i, C_i :

$$A = \sum_{i=1}^k w_i A_i, B = \sum_{i=1}^k w_i B_i, C = \sum_{i=1}^k w_i C_i \quad (4)$$

For the realization of the linear controller, the controllability of all local fuzzy models is required. For a magnetic levitation system, this condition is satisfied. In this case, it is possible to realize stabilizing state feedback controllers for all local models.

Thus, the control is given by formula (5)

$$u = Kx, K = \sum_{i=1}^k w_i K_i \quad (5)$$

In the case of a fuzzy logic controller designed using linear local models, the asymptotical stability of the system is required. Analysis based on the Lyapunov stability theorem [24–26] gives satisfactory results.

4 Active Magnetic Levitation

The magnetic levitation phenomenon allows us to suspend a ferromagnetic object without contact with the surrounding field [9]. Being an example of a nonlinear and unstable system, magnetic levitation is an ideal tool for teaching and research purposes. Typically, controlled magnetic levitation systems are based on a single electromagnet located at the top of the frame. The electromagnetic force generated by the electromagnet counteracts the gravity force, so the object can levitate. The controlled magnetic levitation is mostly based on a distance sensor. A controller implemented in analogue and/or digital [22] form controls the current in the electromagnet coil. In the research, the MLS1EM system was modified with a hardware current controller based on chopping technique. Controlling the system with MATLAB/Simulink operating in real-time mode on NI PCI 6251 board with a sampling frequency of 1kHz, it was possible to neglect

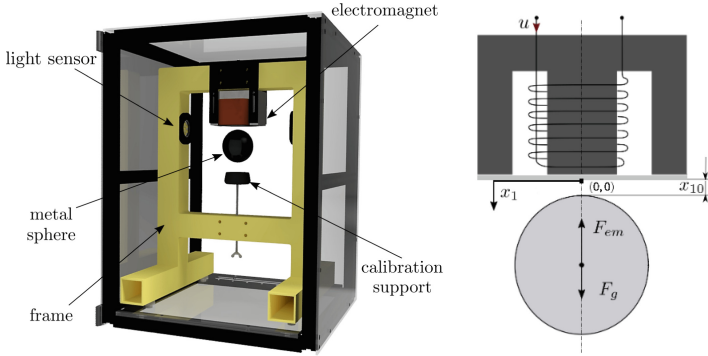


Fig. 2. MLS1EM active magnetic levitation test-rig and system diagram.

the actuator dynamics and consider the magnetic levitation system as a current-controlled one. Therefore, it can be described by a differential equation of the form:

$$\begin{cases} \dot{x}_1 = x_2 \\ \dot{x}_2 = \frac{1}{2m}L^{(1)}(x_1) * i^2 + g \end{cases} \quad (6)$$

where: x_1 - armature distance to the electromagnet [m], ($x_{10} \in (0, x_{1max})x_1 > 0$) - position of levitation, x_2 - armature velocity [m/s], m - armature mass [kg], g - gravitational acceleration [m/s²], $L(.)$ - coil inductance [H], i - current in the coil of the electromagnet [A] ($i \geq 0$) (Fig. 2).

Parameter	Unit	Value
x_{1min}	[m]	0
x_{1max}	[m]	0.0105
x_{2min}	[m/s]	-2.34
x_{2max}	[m/s]	0.42
m	[kg]	0.03155
g	[m/s ²]	9.81
i_{max}	[A]	2

4.1 Linear Model

To linearize the system given in (6) the requested position x_{10} is defined. Then, the requested steady-state coil current i_0 is calculated to be used together with a state feedback controller at the stabilization point.

$$i_0 = \sqrt{\frac{-2mg}{L^{(1)}(x_{10})}} \quad (7)$$

To obtain a simple tunability of the stiffness and damping of the system, the state feedback controller of the form (8) is proposed

$$i(t) = -K_1 x_1(t) - K_2 x_2(t) \quad (8)$$

Knowing that two control coefficients are available, the dynamic properties of the control feedback loop can be configured as desired. Analyzing the characteristic equation of a closed-loop system of the form (9) one can notice that the distribution of eigenvalues depends on the selection of controller parameters for the particular operating point.

$$\dot{x} = \begin{bmatrix} 0 & 1 \\ \alpha - \beta K_1 & -\beta K_2 \end{bmatrix} x \quad (9)$$

The characteristic equation of a closed-loop system is of the form:

$$\lambda^2 + \lambda\beta K_2 + \beta K_1 - \alpha = 0 \quad (10)$$

The controller parameter values are determined by the dynamic properties and the stability criterion of the closed-loop system.

$$\Delta_z = \beta K_2^2 - 4(\beta K_1 - \alpha) \quad (11)$$

The number and type of system eigenvalues depend on the determinant (11) of the characteristic equation (10) of the closed loop system (9).

The permissible values of K_1 and K_2 are limited by the asymptotic stability criterion of the closed-loop system. For the system (9), the following cases are feasible:

$$\begin{aligned} \Delta_z = 0, & K_1 = 0.25\beta K_2^2 + \alpha\beta^{-1} \\ & \lambda_1 = \lambda_2 = -0.5\beta K_2 \\ \Delta_z > 0, & K_1 < 0.25\beta K_2^2 + \alpha\beta^{-1} \\ & \lambda_{1,2} = -0.5(\beta K_2 \pm \sqrt{\Delta_z}) \\ \Delta_z < 0, & K_1 > 0.25\beta K_2^2 + \alpha\beta^{-1} \\ & \lambda_{1,2} = -0.5(\beta K_2 \pm j\sqrt{-\Delta_z}) \end{aligned} \quad (12)$$

The asymptotic stability condition $Re(\lambda_i) < 0$, $i = 1, 2$ is satisfied for $-\beta K_2 < 0$, hence $K_2 < 0$. The feedback parameters should be determined so that the damping and elastic coefficients of the closed system are given in advance (Figs. 4 and 5).

It can be seen that the linear system is observable and controllable. The eigenvalues λ_1 and λ_2 depending on the operating point are presented in Fig. 3a. One can find the structural instability of the system due to the positive sign of one of the eigenvalues. The steady-state current (i_0) versus the sphere positions (x_{10}) is presented in Fig. 3b. To satisfy stable operation, the control law is given

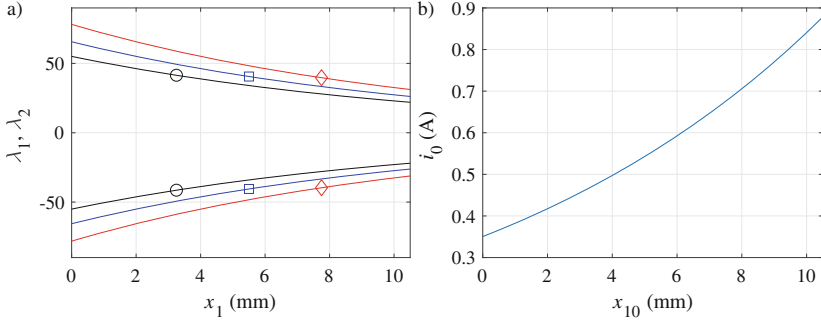


Fig. 3. Eigen values distribution and coil current vs ball position

Table 1. Controller parameters for the aperiodic property set for all steady-state points

$x_{10}[mm]$	3.25	5.25	7.25
$K_1[A/m]$	-174.42	-204.50	-239.78
$K_2[As/m]$	-3.64	-4.26	-5
$i_0[A]$	0.372	0.5	0.602

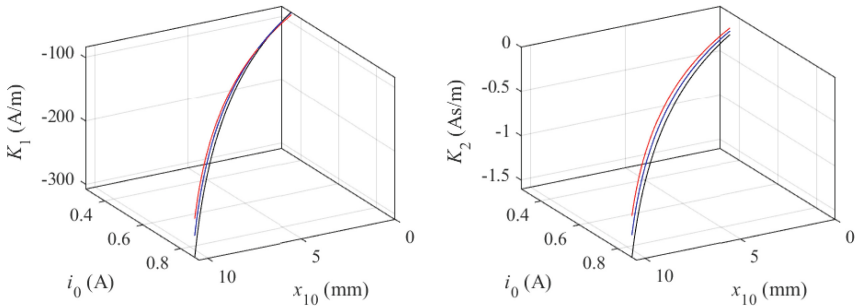


Fig. 4. Controller parameter K_1 and K_2

by Eq. (8). It can be observed that the linearized magnetic levitation system under state feedback control corresponds to the ideal mechanical equivalent with a programmable form of elasticity and damping coefficients. The configurable dynamics means that the controller parameters are calculated on the basis of the requested eigenvalue distribution. Two cases were considered: when they were fixed at all three considered stabilization positions (aperiodic) (cf. Table 1), and when they were different in the central position (oscillatory) (cf. Table 2).

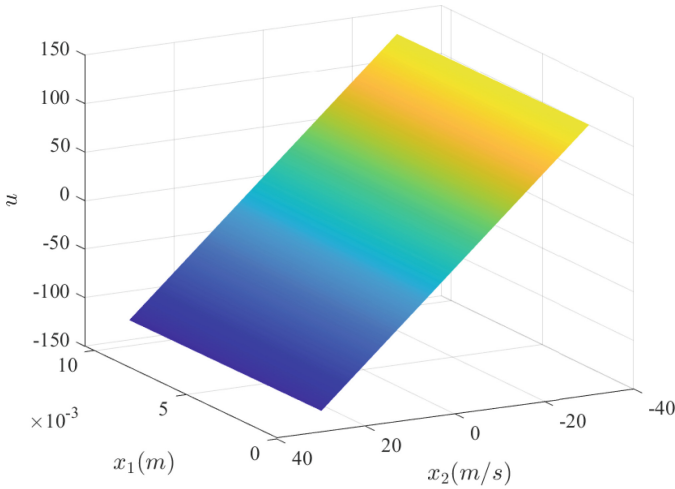


Fig. 5. Output of the controller with respect to the system state: displacement and velocity

Table 2. Controller parameters for the aperiodic property set external steady-state points and oscillatory for the middle point

$x_{10}[mm]$	3.25	5.25	7.25
$K_1[A/m]$	-174.42	-134.63	-239.78
$K_2[As/m]$	-3.64	-0.54	-5
$i_0[A]$	0.372	0.49	0.602

4.2 Fuzzy Logic State Feedback Controller

As part of the research, the functionality of the FLTune toolbox developed [27] was extended by a method of automatic generation of the structure of the fuzzy controller with the output constituting the controller settings of the state and control in a steady state on the basis of set eigenvalues or parameters determining the elasticity and damping of the magnetic suspension. In this way, a universal tool was obtained to test various configurations of fuzzy controllers synthesized on the basis of linear models. Four configurations of the fuzzy controller were tested, in which two types of membership functions were established: triangular and Gaussian. Tangent and overlapping membership functions were used (see Fig. 6). In the case of overlapping triangular membership functions, a linear transition can be obtained between the determined output values from the fuzzy structure. Separation of the membership function guarantees the constancy of outputs in specific intervals of the input argument. The use of Gaussian membership functions makes it possible to obtain a nonlinearity in the switching output.

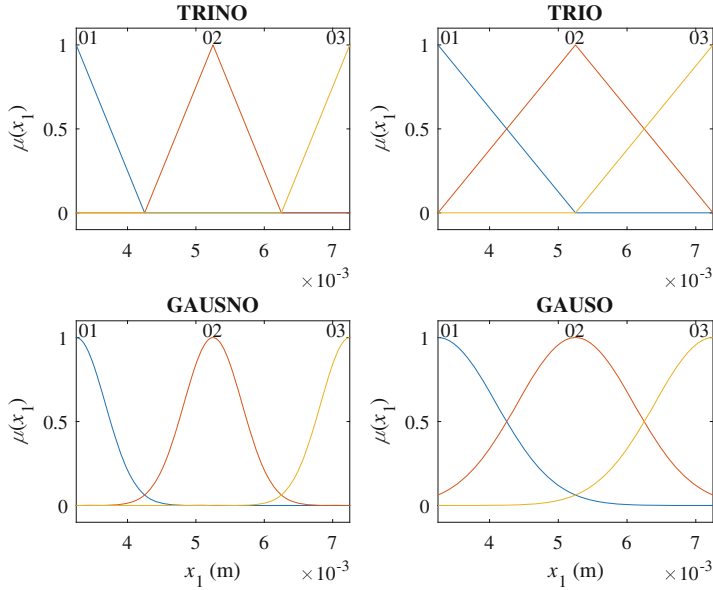


Fig. 6. Four considered configurations of membership functions configurations defined for ball position input

5 Validation by Sine-Wave Tracking

All four scenarios were tested in a sine-wave tracking task. In Fig. 8 the FLC variable output and in Fig. 9 summarized control are presented in the simulation stage. The displacement errors obtained for a selected type of FLC are given in Fig. 10. Stabilization errors vary with respect to the method considered. Unsurprisingly, the best result was obtained for the fuzzy logic controller with overlapped triangular membership functions. In general, the positioning error is in the range of $\pm 300 \mu\text{m}$. Therefore, for the tracking problem, the FLC with overlapping MFs is requested to be applied. Otherwise, non-overlapping MFs can only be used for the configured steady-state point (Fig. 7).

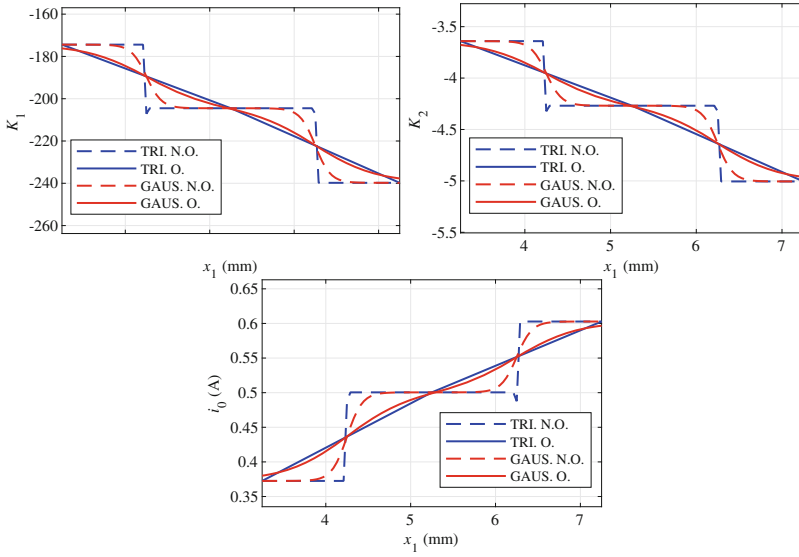


Fig. 7. Controller gain K_1 , K_2 and steady-state current i_0

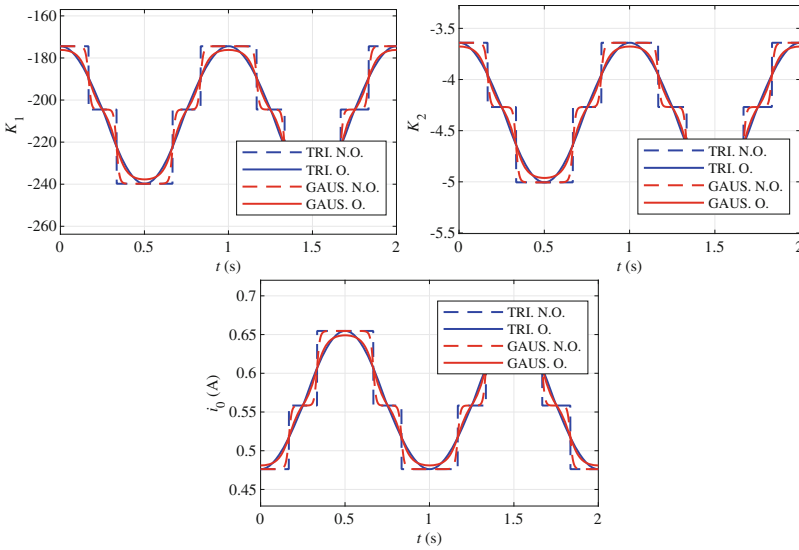


Fig. 8. Controller gain K_1 , K_2 and steady-state current i_0 for sine wave tracking

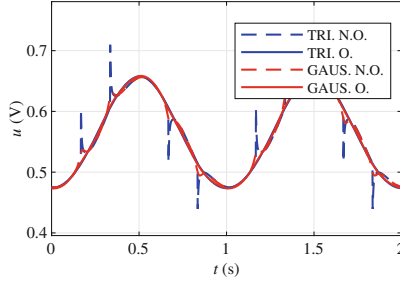


Fig. 9. Control signals

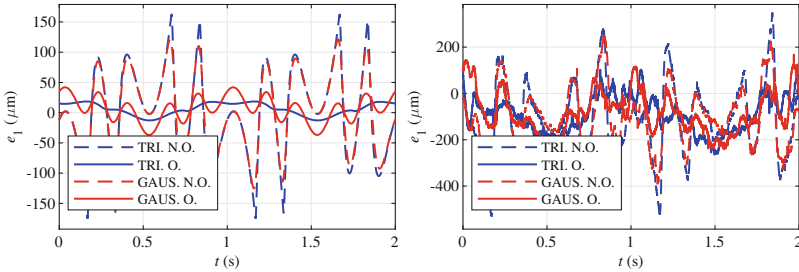


Fig. 10. Displacement error for sine wave tracking mode - simulation and experiment

6 Research on Two Closed-Loop Dynamics Configurations

The electromagnetic suspension is a very specific system in the case of levitated object motion. It is requested that the upper and position in the whole operating range can be accessed without any overshoots. Therefore, the response of the system must be aperiodic or aperiodic critical. In the remaining range the overshoots can be acceptable depending on their amplitude. Moreover, one can imagine that the oscillatory mode is requested at some operating distance. To carry out the levitation tests of the object with the designed dynamic properties for different positions of the sphere, a test signal of the set point was developed according to the following scenario:

$$x_{1d} = \begin{cases} x_{10}, & 0 \leq t < t_1 \\ x_{10} + dx, & t_1 \leq t < t_2 \\ x_{10}, & t_2 \leq t < t_3 \\ x_{10} - dx, & t_3 \leq t < t_4 \\ x_{10}, & t_4 \leq t < t_f \end{cases} \quad (13)$$

The interval time sequence was configured in the supervisory control system, designing the experiment to be carried out at t_f . The positioning of the sphere was performed with $dx = 2$ mm around the operating point $x_{10} = 5.25$ mm according to the designed controller configuration (Fig. 6).

Research on configuration with fixed eigenvalues. In the first scenario, the eigenvalues were set to (-75) for all desired positions. One can find that the sphere is stabilised with the requested performance (see Fig. 11).

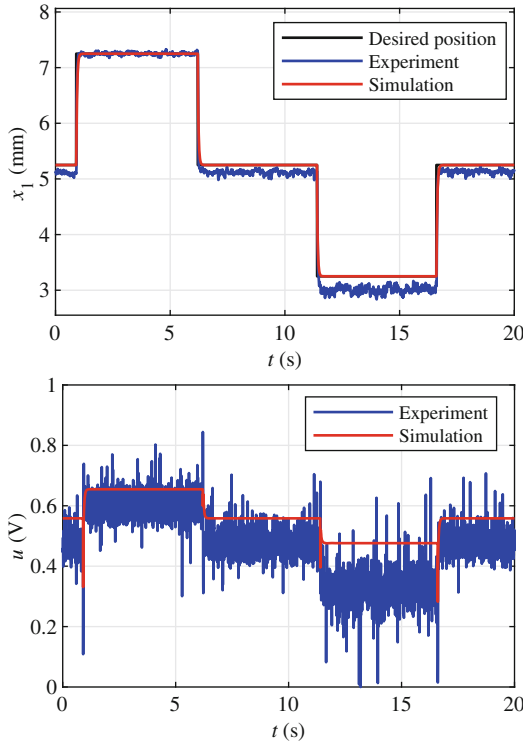


Fig. 11. Sphere stabilisation at fixed eigenvalues set controller design: displacement and control signals vs time

Research on configuration with variable eigenvalues. In the second scenario, the eigenvalues were set to (-75) for external positions and $(-9.51 \pm j55.49)$ for the middle desired position. One can find that the sphere is stabilized and the dynamics of the closed-loop system is modified as requested (see Fig. 12).

The simulation results confirm the study and the requested control action. There is a mismatch between simulation and experiment, which shows that the model needs to be upgraded.

From the experimental data, information about the current controller operation can be extracted. Its variable frequency, and therefore the consequences for stabilization, are well visible. Moreover, the motion up and down causes the lateral vibrations to be uncontrolled, and therefore the lateral motion [28] is observed in the axial position measurements, which affects the control quality. One can find differences between the model and the experiment. This method

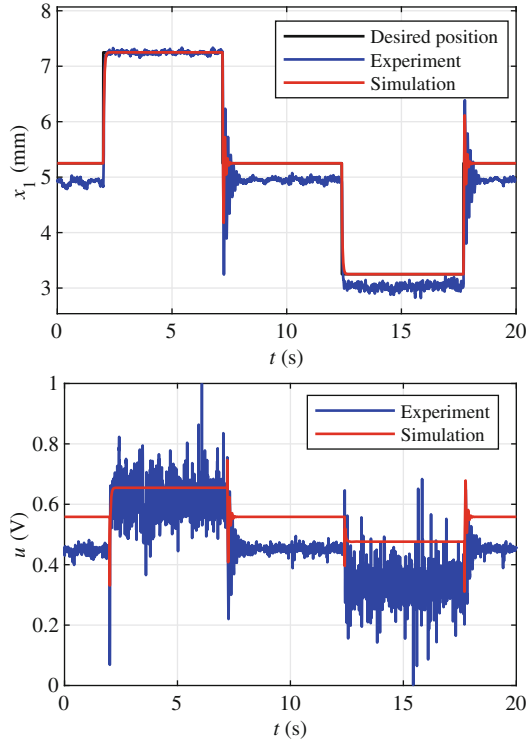


Fig. 12. Sphere stabilisation at variable eigenvalues set for controller design: displacement and control signals vs time

can be used to tune up the identified model to obtain a full convergence, as well as for online adaptation of the fuzzy logic controller to obtain the requested closed-loop dynamics.

7 Conclusions

The implementation of the research contributed to the development of the FLTune Toolbox, within which the methods of automatic generation of the fuzzy controller configuration were extended and a method of conducting research in the real-time regime with the use of supervisory control with the possibility of data archiving, changing the set position of the levitating object and selecting the controller was developed. In this work, the possibilities of designing a control system for a dynamic system to obtain the desired dynamic properties as a function of the position of the levitating object were investigated and demonstrated. The obtained results for both continuous linear and nonlinear changes in properties as well as interval changes constitute an important solution for further applications in the field of axial magnetic bearing control [29], stabilization and

design of the two-legged [30] gait dynamics and control of the cooling system in the Data Center when switching between free-cooling and compressor [31] operating modes. Further research work will be oriented towards extending the study by using other state variables to obtain the MISO regulator.

References

1. Ali Awad, O., Laith Salim, I.: Fuzzy PID gain scheduling controller for networked control system. *Iraqi J. Sci.* 210–216 (2021) 1
2. Kamala, N., Thyagarajan, T., Renganathan, S.: Fuzzy gain scheduled multivariable control of nonlinear system using PSO based PID. 403–408, 3892–3899 (2012)
3. Tran, H.K., Lam, P.D., Trang, T.T., Nguyen, X.T., Nguyen, H.N.: Fuzzy gain scheduling control apply to an RC hovercraft. *Int. J. Electr. Comput. Eng.* **10**, 2434–2440 (2020) 8
4. Zhao, Z.Y., Tomizuka, M., Isaka, S.: Fuzzy gain scheduling of PID controllers. *IEEE Trans. Syst. Man Cybernet.* **23**, 1392–1398 (1993)
5. Hady, F., Abuelenin, S.: Design and simulation of a fuzzy-supervised PID controller for a magnetic levitation system. *Studies Inform. Control* **17**, (2008) 10
6. Unni, A.C., Junghare, A.S., Mohan, V., Ongsakul, W.: PID, fuzzy and LQR controllers for magnetic levitation system. Institute of Electrical and Electronics Engineers Inc., (2016) 11
7. Yadav, S., Tiwari, J.P., Nagar, S.K.: Digital control of magnetic levitation system using fuzzy logic controller (2012)
8. Pilat, A., Turnau, A.: Self-organizing fuzzy controller for magnetic levitation system. In: *Computer Methods and Systems*, CMS 14–16 November, Kraków, Poland, 2005, pp. 101–106 (2005)
9. Pilat, A.: Control of magnetic levitation systems. PhD thesis, AGH, Krakow (2002)
10. Takagi, T., Sugeno, M.: Fuzzy identification of systems and its applications to modeling and control. *IEEE Trans. Syst. Man Cybernet. SMC* **15**, 116–132 (1985)
11. De Freitas Nunes, E.A., et al.: Proposal of a fuzzy controller for radial position in a bearingless induction motor. *IEEE Access* **7**, 114808–114816 (2019)
12. Ahmad, A.K., Saad, Z., Osman, M.K., Isa, I.S., Sadimin, S., Abdullah, S.S.: Control of magnetic levitation system using fuzzy logic control. In: *Proceeding - 2nd International Conference Computational Intelligence, Modelling and Simulation, CIMSIm 2010*, pp. 51–56 (2010)
13. García-Gutiérrez, G., et al.: Fuzzy logic controller parameter optimization using metaheuristic cuckoo search algorithm for a magnetic levitation system. *Appl. Sci.*, **9**(12) (2019)
14. Zhang, J., Wang, X., Shao, X.: Design and real-time implementation of Takagi-Sugeno fuzzy controller for magnetic levitation ball system. *IEEE Access* **8**, 38221–38228 (2020)
15. David, R.-C. Dragos, C.-A., Bulzan, R.-G., Precup, R.-E., Petriu, E.M., Radac, M.-B.: An approach to fuzzy modeling of magnetic levitation systems. Technical Report A12 (2012)
16. Ma, Z., Liu, G., Liu, Y., Yang, Z., Zhu, H.: Research of a six-pole active magnetic bearing system based on a fuzzy active controller. *Electronics* **11**(11) (2022)
17. Minihan, T.P., Lei, S., Sun, G., Palazzolo, A., Kascak, A.F., Calvert, T.: Large motion tracking control for thrust magnetic bearings with fuzzy logic, sliding mode, and direct linearization. *J. Sound Vib.* **263**(3), 549–567 (2003)

18. Santisteban, J.A., Mendes, S.R.A., Sacramento, D.S.: A fuzzy controller for an axial magnetic bearing. In: 2003 IEEE International Symposium on Industrial Electronics (Cat. No.03TH8692), vol. 2, pp. 991–994 (2003)
19. Regaya, C.B., Zaafouri, A., Chaari, A.: A new sliding mode speed observer of electric motor drive based on fuzzy-logic. Technical Report 3
20. Zheng, W., Xia, B., Wang, W., Lai, Y., Wang, M., Wang, H.: State of charge estimation for power lithium-ion battery using a fuzzy logic sliding mode observer. *Energies* **12**(13) (2019)
21. Lin, F.-C., Yang, S.-M.: Adaptive fuzzy logic-based velocity observer for servo motor drives. Technical Report
22. Pilat, A.K.: Active Magnetic Levitation Systems. AGH University of Science and Technology Press, Krakow, Poland (2013)
23. Pilat, A.: Analytical modeling of active magnetic bearing geometry. *Appl. Math. Model.* **34**, (2010)
24. Nguyen, H., Prasad, R.: Fuzzy Modeling and Control: Selected Works of M. CRC Press (1999)
25. Joh, J., Chen, Y.-H., Langari, R.: On the stability issues of linear Takagi-Sugeno fuzzy models. *IEEE Trans. Fuzzy Syst.* **6**, 402–410 (1998)
26. Teixeira, M.C.M., Zak, S.H.: Stabilizing controller design for uncertain nonlinear systems using fuzzy models. *IEEE Trans. Fuzzy Syst.* **7**(2), 133–142 (1999)
27. Pilat, A.K.: Fltune - automatic fuzzy logic controler desing toolbox. Technical Report, AGH University of Science and Technology, Department of Automatic Control and Robotics (2010)
28. Pilat, A.K., Sikora, B., Zrebiec, J.: Investigation of lateral stiffness and damping in levitation system with opposite electromagnets*. In: 2019 12th Asian Control Conference (ASCC), pp. 1210–1215 (2019)
29. Sikora, B.M., Pilat, A.K.: Analytical modeling and experimental validation of the six pole axial active magnetic bearing. *Appl. Math. Model.* **104**, 50–66 (2022)
30. Milanowski, H., Pilat, A.K.: Comparison of identified and simscape model of human leg motion (2020) 09
31. Borkowski, M., Pilat, A.K.: Customized data center cooling system operating at significant outdoor temperature fluctuations. *Appl. Energy* **306**, 117975 (2022)



Consistent Design of PID Controllers for Time-Delay Plants

Andrzej Bożek^(✉), Zbigniew Świder, and Leszek Trybus

Faculty of Electrical and Computer Engineering, Rzeszów University of Technology,
ul. Wincentego Pola 2, 35-959 Rzeszów, Poland
{abozek, swiderzb, ltrybus}@prz.edu.pl

Abstract. Inspired by IMC approach, tuning rules for PID control of time-delay plants such as pure TD, FOPTD, SOPTD and IPTD are developed. Control systems for the first three plants can be designed in a consistent way, since they are described by the same transfer function for commonly used controllers. Analytic rules are given for critical damping. Nomograms characterize overdamping for settling time specification and underdamping for overshoot. Practical experiment confirms fairly good correspondence of closed-loop responses to specifications.

Keywords: Low-order models · Internal Model Control · PID settings · Identification

1 Introduction

Control loops with PID controllers are most common in process industries. Internal Model Control (IMC) based on direct synthesis is one of design methods used to calculate settings of such controllers [1, 2]. Single specification data, namely closed-loop time constant often denoted by λ , is the essential advantage of IMC from practical viewpoint. In the case of plants with time-delay, after Taylor or Padé approximation and cancellation of time constants, one obtains simple rules for controller settings expressed in terms of plant parameters and λ . Due to simplicity, the PID-IMC tuning rules have been widely approved in industrial practice [2] (also known as SIMC).

However, selection of appropriate λ to obtain fully satisfying closed-loop response requires some experience. Despite that $\lambda = \tau$ (delay) is recommended in [3] to get “tight” response, $\lambda = 1.5\tau$ for smoother response, and $\lambda = 0.5\tau$ for “more aggressive”, some fine tuning of the controller is often needed. Therefore certain improvement of the rules that would provide closed-loop responses better corresponding to specification is the purpose of this paper. Replacing the time constant λ by settling time as specification is one of the solutions. Gain selection nomograms for overdamped and underdamped responses are another one. Note that tuning rules for time-delay systems given an overshoot were developed in [4] using root locus. Design based on gain and phase margin specification can be found for instance in [5].

Plants considered here include pure time-delay TD, first-order-, second-order- and integrator-plus-time-delay, i.e. FOPTD, SOPTD and IPTD, described by the following transfer functions

$$\begin{aligned} TD &: k_o e^{-\tau s}, & FOPTD &: \frac{k_o}{Ts+1} e^{-\tau s}, \\ SOPTD &: \frac{k_o}{(Ts+1)^2} e^{-\tau s}, & IPTD &: \frac{k_o}{s} e^{-\tau s}. \end{aligned} \quad (1)$$

Note that SOPTD corresponds to FOPTD in terms of the parameter set $\{k_o, T, \tau\}$. Such restricted SOPTD is also beneficial for identification. TD, FOPTD and SOPTD are consistent in the sense that after usage of an appropriate controller, namely I, PI, and PID, respectively, and cancellation of the time constant, the same transfer function characterizes each of the system. This implies common tuning rules and nomograms.

The paper is organized as follows. Tuning rules for critical damping in I + TD, PI + FOPTD and PID + SOPTD systems are developed in the next section using Padé approximation. Overdamping for settling time specification and underdamping for overshoot are considered in Sect. 3, giving easy to use nomograms. Settings for PID + IPTD system with a specified settling time are developed in Sect. 4. The method is practically verified in Sect. 5 using a temperature control system with plant identification. Conclusions are given at the end.

2 Critical Damping

As a reference consider the pure time-delay plant TD of (1) controlled by integral I controller k_I/s with the gain k_I . The open-loop transfer function becomes

$$G_{open}(s) = k \frac{e^{-\tau s}}{s} \quad (2)$$

where $k = k_I k_o$. The same G_{open} is obtained for PI + FOPTD system after time constant cancellation $T_I = T$ in the controller $k_p(T_I s + 1)/(T_I s)$ [1, 2]. Here $k = k_p k_o / T$. Likewise for PID + SOPTD with the controller in the form

$$k_p \left(1 + \frac{1}{T_I s} + T_D s\right) = k_p \frac{\left(\frac{T_I}{2} s + 1\right)^2}{T_I s}, \quad T_D = \frac{T_I}{4} \quad (3)$$

we get (2) for $T_I = 2T$ with $k = k_p k_o / (2T)$. Hence the same closed-loop response will be generated for each of the I + TD, PI + FOPTD and PID + SOPTD systems given an open-loop gain k .

By using the 1st order Padé approximation of $e^{-\tau s}$ we get

$$G_{open}(s) = k \frac{-\frac{\tau}{2} s + 1}{s\left(\frac{\tau}{2} s + 1\right)} \quad (4)$$

or, after time scale normalization,

$$G_{open}(s') = k' \frac{-s' + 1}{s'(s' + 1)}, \quad (5)$$

$$s' = \frac{\tau}{2}s, k' = \frac{\tau}{2}k.$$

The closed-loop transfer function becomes

$$G_{closed}(s') = \frac{k'(-s' + 1)}{s'^2 + (1 - k')s' + k'}. \tag{6}$$

Critical damping requires denominator discriminant equal to zero, what in the case of G_{closed} yields the relative gain k' and the double pole $s'_{1,2}$ as [4]

$$k' = 3 - 2\sqrt{2} \cong 0.17, \quad s'_{1,2} = -(\sqrt{2} - 1) \cong -0.41. \tag{7}$$

Hence the gain k , time constant $T_{closed} = \frac{\tau}{2}/|s'_{1,2}|$ and settling time t_s are given by

$$k = 0.34\frac{1}{\tau}, \quad T_{closed} = 1.21\tau, \quad t_s = 6T_{closed} = 7.26\tau \tag{8}$$

(98% of steady-state output). Using the earlier expressions for k in the case of I, PI and PID controllers we get the settings given in Table 1. Note that T_{closed} determined in (8) is a little more than τ recommended in [3] for a tight response. One can check that the closed-loop responses obtained for a plant with $e^{-\tau s}$ and with Padé approximation almost overlap except for the very beginning.

Table 1. Controller settings for critical damping

I + TD	PI + FOPTD	PID + SOPTD
$k_I = 0.34\frac{1}{k_o\tau}$	$k_P = 0.34\frac{1}{k_o}\frac{T}{\tau}$ $T_I = T$	$k_P = 0.68\frac{1}{k_o}\frac{T}{\tau}$ $T_I = 2T, T_D = T/2$

3 Overdamping and Underdamping

There are sensitive processes where slower operation of the controller than the one before is preferred. For k' smaller than 0.17, the denominator in (6) has two distinct real roots. Since development of an analytic expression for the settling time would be too involved, we have turned to simulations. It does not matter whether $e^{-\tau s}$ or Padé approximation is used since the responses, except for the beginning, almost overlap.

The nomogram in Fig. 1 allows to get the relative gain $k' = \frac{\tau}{2}k$ given the relative settling time t_s/τ . For instance, if $t_s/\tau = 15$, i.e. twice longer than for critical damping in (8), then $k' = 0.107$. Having $k = k'\frac{2}{\tau}$, controller settings are obtained from corresponding expressions in the previous section.

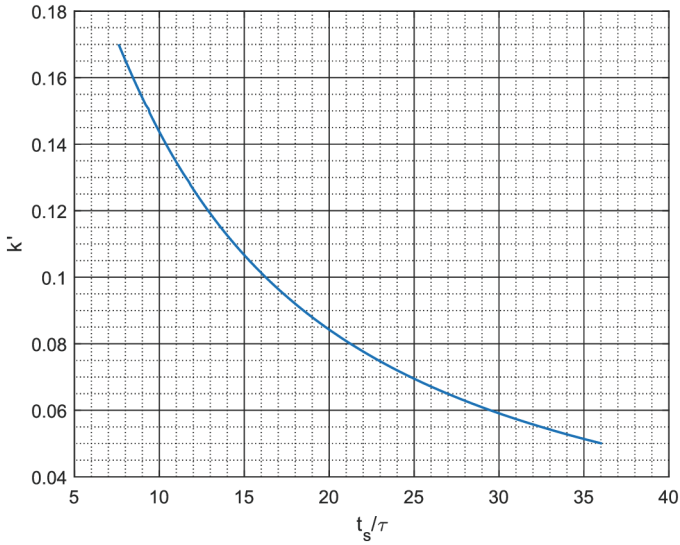


Fig. 1. Relative gain in terms of settling time

Small overshoot of about 5% is often required in practice as it reveals that the system behavior is close to critical damping. Larger overshoots are encountered rarely, only to improve disturbance suppression.

For $k' > 0.17$ the denominator in (6) has complex poles, so formulae for the percentage overshoot $OVS_{\%}$ and settling time could be developed using root locus, as shown in [4]. However, contrary to the critically- and overdamping cases, the closed-loop responses for $e^{-\tau s}$ or Padé approximation are now somewhat different, so to get precise results, we have again turned to simulations involving $e^{-\tau s}$.

The nomogram in Fig. 2 determines the relative gain k' for a given overshoot $OVS_{\%}$. For $OVS_{\%} = 5$ the nomogram gives $k' = 0.258$. Then the open-loop gain k and controller settings are obtained as above.

For the overshoot $OVS_{\%}$ from 15 up to over 27 the settling time t_s is close to the one in (8). Smaller $OVS_{\%}$ corresponds to a shorter t_s , i.e. about 6τ . $OVS_{\%} = 30$ requires $t_s \cong 9\tau$.

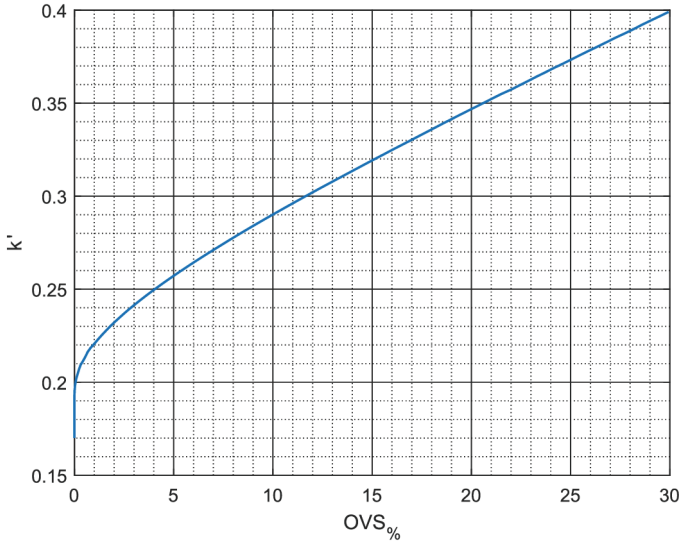


Fig. 2. Relative gain in terms of the overshoot

4 Control of IPTD

Last component of a multi-stage process usually accumulates the final product, so it may be described by an integrator with time-delay, i.e. IPTD in (1). The PID controller in serial form is applied here, so

$$k_P \frac{(T_1 s + 1)(T_2 s + 1)}{T_I s}, T_1 > T_2, T_I = T_1 + T_2, T_D = \frac{T_1 T_2}{T_I} \leq \frac{T_I}{4}. \tag{9}$$

Replacing $e^{-\tau s}$ by Padé approximation and taking $T_2 = \frac{\tau}{2}$ (cancellation) gives the transfer function

$$G_{open}(s) \cong \frac{k_P k_o}{T_I} \frac{(T_1 s + 1)(-\frac{\tau}{2} s + 1)}{s^2}. \tag{10}$$

Using $s' = \frac{\tau}{2} s$ yields

$$G_{closed}(s) \cong \frac{k'(T_1' s' + 1)(-s' + 1)}{(1 - k' T_1') s'^2 + k'(T_1' - 1) s' + k'}, \tag{11}$$

$$k' = \frac{1}{4} \frac{k_P k_o}{T_I} \tau^2, \quad T_1' = T_1 \frac{2}{\tau}.$$

The conditions $k' < 1/T_1'$ and $T_1' > 1$ provide stability. However, T_1' should be at least a few (say, 5) to obtain almost overlapping responses for $e^{-\tau s}$ and Padé.

To make the development more clear we introduce an intermediate design coefficient

$$\eta = T_1', \quad \text{so} \quad T_1 = \eta \frac{\tau}{2}. \tag{12}$$

Since T_1 is a time parameter of PID, so value of η affects the settling time t_s and vice versa.

To get analytical results we restrict considerations to zero discriminant of the denominator (critical damping) $(1 - k'\eta)s^2 + k'(\eta - 1)s' + k'$. This is characterized by

$$k' = \frac{4}{(\eta + 1)^2}, \quad s'_{1,2} = -\frac{2}{\eta - 1}. \tag{13}$$

Since $t'_s = 6/|s'_{1,2}|$, so $t_s = 3(\eta - 1)\frac{\tau}{2}$. Assuming that the settling time is a design data we obtain

$$\eta = \frac{2}{3} \frac{t_s}{\tau} + 1 \tag{14}$$

as the starting point for calculation of PID settings given in Table 2. T_I and T_D follow from (9) for $T_1 = \eta\frac{\tau}{2}$ and $T_2 = \frac{\tau}{2}$. Using k' from (13) in (11) gives k_P .

Table 2. PID settings for IPTD process

η	k_P	T_I	T_D
$\frac{2}{3} \frac{t_s}{\tau} + 1$	$\frac{8}{\eta+1} \frac{1}{k_o\tau}$	$(\eta + 1)\frac{\tau}{2}$	$\frac{\eta}{\eta+1} \frac{\tau}{2}$

Sample data $k_o = 1, \tau = 2, t_s = 12(= 6\tau)$ yield $\eta = 5, T_1 = 5, T_2 = 1, T_I = 6, T_D = 5/6$ and $k_P = 2/3$. Note that the set-point filter $1/(T_1s + 1)$ is needed in the system to eliminate the corresponding component in the numerator of (10) thus avoiding an overshoot. Usage of 2DOF-PID controller can be an alternative [6, 7].

5 Practical Verification

A simple lab setup shown in Fig. 3 has been assembled to verify the design method. The setup involves a heating resistor kept in open air, Pt100 temperature sensor, transducer, PWM switch, and an embedded system communicating with PC running Matlab. Roughly similar assembly has been described in [8]. The embedded system records the measurements, works as PID controller, generates a PWM signal, and supervises the whole operation. Ambient temperature drift is treated as a load disturbance. The file recorded by the embedded system is sent to PC for identification.

For the nominal $PWM = 25\%$ the resistor temperature is about 76 °C. After increasing PWM to 27% the temperature increases by over 2.5 °C reaching steady-state in 30 min. Such modest 2% increase of the input may correspond to technological restrictions in an industrial case. The green belt in Fig. 4 presents measurements acquired by Pt100. The measurements are smoothed out by a running average with 1 min window what gives the experimental response shown in blue.

The question now is whether FOPTD or SOPTD model will be more appropriate for the smoothed response. The question may be answered by determining location of the

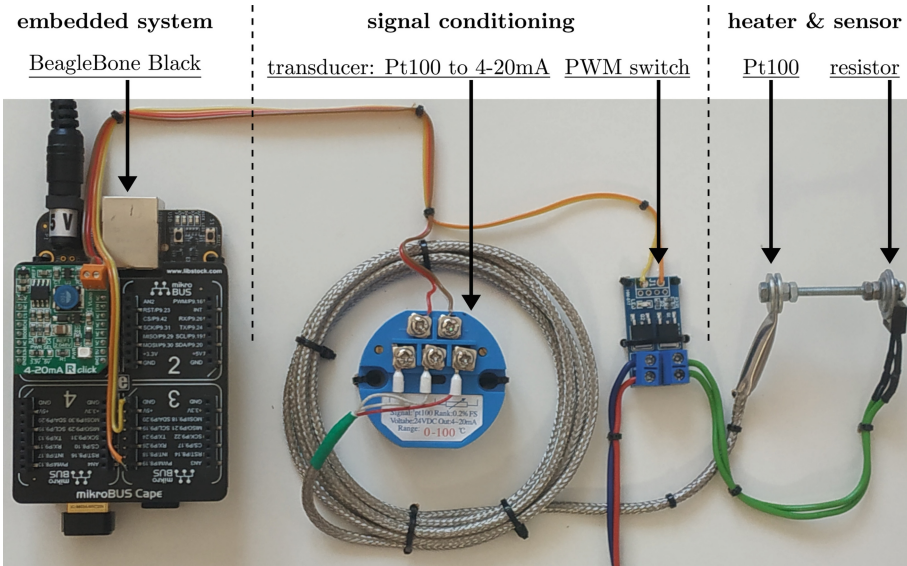


Fig. 3. Lab setup

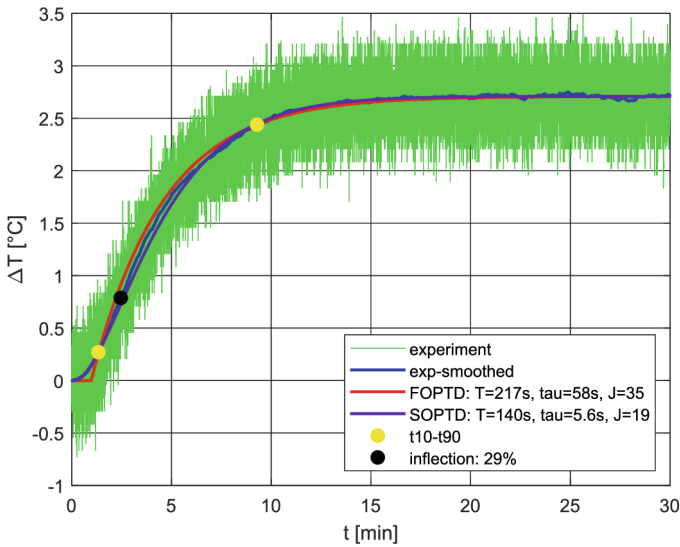


Fig. 4. Step responses of the plant and models

inflection point (black dot in Fig. 4). Note that for SOPTD with double time constant the inflection point is at 26% of the steady-state value. Hence height of the inflection point may determine selection of the model as

FOPTD : below 20%, SOPTD : above 25%.

For locations between 20% and 25% either FOPTD or SOPTD can be chosen. During 10 repetitions of the step test, SOPTD model has been selected more often than FOPTD.

The plant gain k_o is about 1.35 [°C/PWM%]. Parameters T , τ of the selected model can be determined by choosing two representative points in the response, first below, the other above the inflection. Let t_{10} , t_{90} denote the times at which the response reaches 10% and 90%, respectively (yellow dots in Fig. 6). Selection of other representative points is reviewed in [9] (see also [10]). Given t_{10} , t_{90} the parameters are calculated as [11]

$$\text{FOPTD} : T = (t_{90} - t_{10})/2.2, \tau = t_{10} - 0.10T \quad (15)$$

$$\text{SOPTD} : T = (t_{90} - t_{10})/3.3, \tau = t_{10} - 0.53T.$$

Average values from the repetitions are: FOPTD $T = 210$, $\tau = 60$; SOPTD $T = 120$, $\tau = 10$ (in seconds). In general, SOPTD better approximated the responses what is seen from the least-squares errors J in the legend (Fig. 4). However, the relative standard deviation of the identified τ in the repetitions is considerably larger for SOPTD than for FOPTD, whereas the deviation for T is similar. It disturbs repeatability of settings and responses of the PID + SOPTD system, since the controller gain depends on T/τ (see Sect. 2). Therefore the FOPTD model is preferred here.

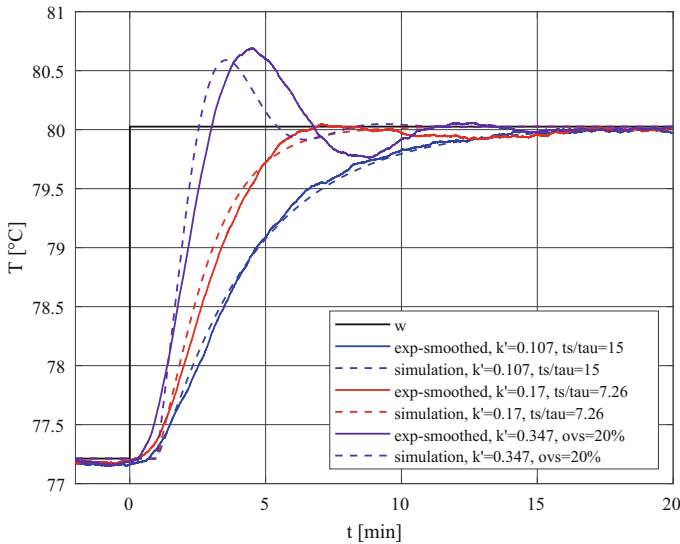


Fig. 5. Closed-loop responses for critical damping, overdamping and underdamping

Closed-loop step responses for the reference change to 80 °C are shown in Fig. 5 for the PI + FOPTD system. Continuous lines present responses from the lab setup, whereas the dotted ones – simulations. Red lines refer to critical damping ($k' = 0.17$) with controller settings from Table 1 and the expected settling time $t_s = 7.26$ min

($\tau = 60$ s). $t_s/15$ is a design data for overdamping, so $t_s = 15$ min. The relative gain $k' = 0.107$ follows from the nomogram in Fig. 1. Likewise, $OVS = 20\%$ is a design data for underdamping, with $k' = 0.347$ read out from Fig. 2. As seen, the experimental and simulated responses for critical damping and overdamping are fairly close, whereas for underdamping they differ somewhat (depending on overshoot).

6 Conclusions

Following Padé approximation and time constant cancellation from IMC design, tuning rules for I + TD, PI + FOPTD, PID + SOPTD and PID + IPTD control systems are developed for critical damping, overdamping and underdamping. The standard PID-IMC rules do not allow for such distinction.

To compare the two approaches note that controller tuning is needed in process automation in two cases:

- initial basic tuning to satisfy specification
- retuning after some time due to process parameter changes.

The rules developed here, suitable for the basic tuning, provide expected closed-loop behavior. Repeatability of that behavior is the objective of the retuning. For this case the IMC rules [1, 2] with the closed-loop time constant λ obtained from the basic tuning are more appropriate due to smaller sensitivity of controller settings to parameter changes.

References

1. Skogestad, S.: Simple analytic rules for model reduction and PID controller tuning. *J. Process Control* **13**(4), 291–319 (2003)
2. Seborg, D.E., Edgar, T.F., Mellichamp, D.A., Doyle, F.J.: *Process Dynamics and Control*, 4th edn. Wiley, New York (2016)
3. Grimholt, C., Skogestad, S.: Optimal PI and PID control of first-order plus delay processes and evaluation of the original and improved SIMC rules. *J. Process Control* **70**, 36–46 (2018)
4. Trybus, L.: A set of PID tuning rules. *Arch. Control Sci.* **15**(1), 5–17 (2005)
5. Hu, W., Xiao, G., Li, X.: An analytical method for PID controller tuning with specified gain and phase margins for integral plus time delay processes. *ISA Trans.* **50**(2), 268–276 (2011)
6. Åstrom, K.J., Hägglund, T.: *Advanced PID control*, Research Triangle Park (2005)
7. Viteckova, M., Vitecek, A.: Two-degree of freedom controller tuning for integral plus time delay plants. *ICIC Express Letters* **2**(3), 225–229 (2008)
8. Oliveira, P.M., Hedengren, J.D.: An APMonitor temperature lab PID control experiment for ungraduated students. In: 2019 24th IEEE International Conference on Emerging Technologies and Factory Automation (EFTA). IEEE (2019)
9. Liu, T., Wang, Q.-G., Huang, H.-P.: A tutorial review on process identification from step or relay feedback test. *J. Process Control* **23**(10), 1597–1623 (2013)
10. Bielińska, A.: Classical methods of step response identification. In: Kasprzyk, J. (ed.) *Process Identification: Joint Publications*, Wyd. PŚI, Gliwice, pp. 51–62 (1997). (in Polish)
11. Trybus, L., Bożek, A.: On feasibility of tuning and testing control loops by nonstandard inputs. In: Bartoszewicz, A., Kabzinski, J., Kacprzyk, J. (eds.) *Advanced, Contemporary Control*. AISC, vol. 1196, pp. 678–688. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-50936-1_57



Synchronization of Four Different Chaotic Communication Systems with the Aim of Secure Communication

Ali Soltani Sharif Abadi¹(✉), Pooyan Alinaghi Hosseinabadi², and Andrew Ordys¹

¹ Institute of Automatic Control and Robotics, Faculty of Mechatronics,
Warsaw University of Technology, Warsaw, Poland
ali.soltani_sharif_abadi.dokt@pw.edu.pl

² School of Engineering and Information Technology, The University of New South Wales,
Canberra, ACT, Australia

Abstract. The synchronization of systems with the goal of securing communications is a very interesting and recent topic. This paper proposes the Fast Terminal Sliding Mode Control (FTSMC) method to design control inputs for synchronizing four different communication systems, Volta, Couillet, Chen and A.M.Rucklidge. Synchronization of these systems can be used in a variety of applications. Some of the key features of the proposed designs are that these inputs are robust against the variety of uncertainties of the system model and external disturbances. Also, the finite time stability of these systems is guaranteed by using designed inputs. Additionally, these inputs eliminate destructive and undesirable chattering phenomena thoroughly.

Keywords: SMC · synchronization · secure communication · chaotic systems · control

1 Introduction

Secure communications are one of the most interesting areas that have recently attracted the great interest of researchers. Different algorithms and methods have been proposed to create a secure gap for data transmission and receiving so far [1]. They are modulated and sent in different ways for data transmission, then demodulated in the receiver. If the goal is to transmit the data securely, the data content should not be available during the transmission process. To accomplish this goal, the synchronization technique provides a secure way because with this method, even if the data is carved, it is no longer accessible from the contents of the data [2].

Chaotic systems reveal unpredictable and controllable behaviors in different states and conditions, which can be seen in many physical phenomena. One of the important features of chaotic systems is that when they display their chaotic behaviors, they behave much like noise and randomize function; nonetheless, when their parameters are plotted against each other, they show a typical behavior [3]. Much research has been done on chaotic systems, as well as the chaotic state of some physical systems has been obtained

by numerical research. New chaotic communication systems have been introduced in [4, 5], and chaotic states have been reported.

Fast Terminal Sliding Mode Control (FTSMC) is a control method that leads a system to achieve finite-time stability. The finite time stability means the state variables of the system reach zero in a finite time called “Terminal”. The use of this control method has attracted researchers’ interest recently. This method has been employed to control the ship in [6], and the technique has been used for a general system in [7, 8]. One of the research done by the sliding mode control method on communication systems is presented in [9], in which the fractional-order chaotic systems have been synchronized with the goal of secure transmission and receiving message signals. Also, the sliding mode concept has been employed in [10] to achieve secure communication by synchronization, and an observer for a chaotic system has been designed. The communication systems have been synchronized in [11] using adaptive control. Also, a sliding mode control has been employed to synchronize the satellite formation with communication time delay in [12]. In the design of the controller by the FTSMC method, the convergence rate of parameters to zero would be increased by adjusting the control coefficients, which is why the word “Fast” is used.

In this paper, four different chaotic communication systems are synchronized with the goal of secure communication using the FTSMC method. The designed control inputs are robust against various parameter uncertainties and external disturbances and eliminate the unwanted chattering problem. Furthermore, they guarantee finite-time stability. Another key feature of the designed control inputs is that they provide several control parameters to adjust the system’s convergence speed.

2 Lemmas and Mathematical Rules

Definition 1: The function $sig^a(x)$ with the relationship between the function of the absolute $|x|$ and the function $sgn(x)$ is defined as $sig^a(x) = |x|^a sgn(x)$. The function of the sign is also defined as follows [13, 14]:

$$sgn(x) = \begin{cases} 1 & ; x > 0 \\ 0 & ; x = 0 \\ -1 & ; x < 0 \end{cases} \tag{1}$$

Definition 2: The relation between the two absolute functions and the signum function is as follows [15]:

$$|x| = xsgn(x) \tag{2}$$

Lemma 1: For the nonlinear system $\dot{x} = f(x), f(0) = 0, x \in D \subseteq \mathfrak{R}^n, x(0) = x_0$ and assuming the constants ρ_1 to ρ_5 as: $\rho_1 > 0, \rho_2 > 0, \rho_3 > 1, \rho_4 = 1 - \frac{1}{2\rho_3}, \rho_5 = 1 + \frac{1}{2\rho_3}$ and the Lyapunov function $V(x) : \mathfrak{R}^n \rightarrow \mathfrak{R}^+$ exists as a scalar continuous radially unbounded function such that $\dot{V}(x) \leq -\rho_1 V^{\rho_4}(x) - \rho_2 V^{\rho_5}(x)$. Therefore, the equilibrium point $x = 0$ of this system is globally finite time stable, the state variables of

this system reach zero from every initial condition, and the upper bound of the settling time is as $T \leq \pi \rho_3 (\sqrt{\rho_1 \rho_2})^{-1}$ and after this time, the system state variables are exactly equal to zero [16, 17].

Lemma 2: Considering the constants $a_1, a_2, \dots, a_n \in \mathfrak{R}$ and choosing $0 < q < 2$, we have: $|a_1|^q + |a_2|^q + \dots + |a_n|^q \geq (a_1^2 + a_2^2 + \dots + a_n^2)^{\frac{q}{2}}$ [18].

3 Models' Expression and Goal Description

This article aims to synchronize four communication systems, including four Volta, Couillet, Chen, and A.M.Rucklidge systems. In this section, the dynamical equations of these systems are presented separately, their chaotic figures are displayed, and then they are prepared to design the controller.

Volta system is chosen as the master system, where the dynamical equations are described in (3).

$$\begin{cases} \dot{x}_m = -x_m - a_1 y_m - z_m y_m = f_{1_m} \\ \dot{y}_m = -y_m - b_1 x_m - x_m z_m = f_{2_m} \\ \dot{z}_m = c_1 z_m + x_m y_m = f_{3_m} \end{cases} \quad (3)$$

where x_m, y_m , and z_m are the master system states in x, y, and z coordinates, respectively. In this system, when system parameters are as $a_1 = 5, b_1 = 85, c_1 = 0.5$, the system becomes chaotic [9].

The Couillet system is the first slave system with its dynamic equations as follows:

$$\begin{cases} \dot{x}_{s_1} = y_{s_1} + u_{1_{s_1}} = f_{1_{s_1}} + u_{1_{s_1}} \\ \dot{y}_{s_1} = z_{s_1} + u_{2_{s_1}} = f_{2_{s_1}} + u_{2_{s_1}} \\ \dot{z}_{s_1} = a_2 z_{s_1} + b_2 y_{s_1} + c_2 x_{s_1} - x_{s_1}^3 + u_{3_{s_1}} = f_{3_{s_1}} + u_{3_{s_1}} \end{cases} \quad (4)$$

where x_{s_1}, y_{s_1} , and z_{s_1} are the first slave system states in x, y, and z coordinates. Also, $f_{1_{s_1}} = y_{s_1}, f_{2_{s_1}} = z_{s_1}$ and $f_{3_{s_1}} = a_2 z_{s_1} + b_2 y_{s_1} + c_2 x_{s_1} - x_{s_1}^3$. The above system also has a chaotic behavior when the parameters are as $a_2 = -0.45, b_2 = -1.1, c_2 = 0.8$ [19].

The dynamic equations of the A.M.Rucklidge system are as follows, this system is the second slave system.

$$\begin{cases} \dot{x}_{s_2} = -a_3 x_{s_2} + b_3 y_{s_2} - y_{s_2} z_{s_2} + u_{1_{s_2}} = f_{1_{s_2}} + u_{1_{s_2}} \\ \dot{y}_{s_2} = x_{s_2} + u_{2_{s_2}} = f_{2_{s_2}} + u_{2_{s_2}} \\ \dot{z}_{s_2} = -z_{s_2} + y_{s_2}^2 + u_{3_{s_2}} = f_{3_{s_2}} + u_{3_{s_2}} \end{cases} \quad (5)$$

where x_{s_2}, y_{s_2} , and z_{s_2} are the second slave system states in x, y, and z coordinates. Also, $f_{1_{s_2}} = -a_3 x_{s_2} + b_3 y_{s_2} - y_{s_2} z_{s_2}, f_{2_{s_2}} = x_{s_2}$ and $f_{3_{s_2}} = -z_{s_2} + y_{s_2}^2$. The A.M.Rucklidge system is in a chaotic condition when parameters are as $a_3 = 2, b_3 = 6.7$ [20].

The Chen system is the third slave system for which its dynamical equations are given in (6).

$$\begin{cases} \dot{x}_{s_3} = -a_4 x_{s_3} + a_4 y_{s_2} + u_{1_{s_3}} = f_{1_{s_3}} + u_{1_{s_3}} \\ \dot{y}_{s_3} = -x_{s_3} z_{s_3} - (c_4 - a_4) x_{s_3} + c_4 y_{s_3} + u_{2_{s_3}} = f_{2_{s_3}} + u_{2_{s_3}} \\ \dot{z}_{s_3} = x_{s_3} y_{s_3} - b_4 z_{s_3} + u_{3_{s_3}} = f_{3_{s_3}} + u_{3_{s_3}} \end{cases} \quad (6)$$

where x_{s_3}, y_{s_3} , and z_{s_3} are the third slave system states in x, y, and z coordinates. Also, $f_{1_{s_3}} = -a_4x_{s_3} + a_4y_{s_2}, f_{2_{s_3}} = -x_{s_3}z_{s_3} - (c_4 - a_4)x_{s_3} + c_4y_{s_3}$ and $f_{3_{s_3}} = x_{s_3}y_{s_3} - b_4z_{s_3}$. In this system, choosing the parameters $a_4 = 35, b_4 = 3, c_4 = 34.5$ [21].

To achieve the synchronization goal, for each slave system, an error dynamic is formed concerning the master system. Then a suitable control input is designed to ensure the state variables of the error dynamic in a finite time reach zero. Consequently, the state variables of slave systems reach to state variables of the master system.

To determine the error dynamics for each slave system than the master system, the errors are considered as $e_{1_{s_i}} = x_{s_i} - x_m, e_{2_{s_i}} = y_{s_i} - y_m$ and $e_{3_{s_i}} = z_{s_i} - z_m$, and in consequence, the error dynamics will be equal to $i = 1, 2, 3$.

$$\begin{cases} \dot{e}_{1_{s_i}} = f_{1_{s_i}} - f_{1_m} + u_{1_{s_i}} \\ \dot{e}_{2_{s_i}} = f_{2_{s_i}} - f_{2_m} + u_{2_{s_i}} \\ \dot{e}_{3_{s_i}} = f_{3_{s_i}} - f_{3_m} + u_{3_{s_i}} \end{cases} \tag{7}$$

It is enough to design inputs that can stabilize these error dynamics in a finite time. As a result, all the state variables of the slave systems converge to the state variables of the master system in a finite time.

4 Design of Control Inputs and Proof of Stability

According to the error dynamics model presented in (7), in this section, control inputs are designed and proofed to guarantee the system’s globally finite time stability and obliterate the Chattering phenomenon.

Theorem: Assuming the presented system in (7) and the sliding surface and the control input, which are given in (8) and (10), respectively, the globally finite time stability of the error model system, is proven according to Lemma 1. Also, the upper bound of the settling time (stability time) of the system is as $T_s \leq \pi r_3 (\sqrt{r_1 r_2})^{-1}$, where r_3 is a constant greater than one and $r_1 = (\sqrt{2})^{\alpha_{1hi}+1} k_{1hi}, r_2 = (\sqrt{2})^{\alpha_{2hi}+1} k_{2hi}, r_4 = \alpha_{1hi} = 1 - \frac{1}{r_3}, r_5 = \alpha_{1hi} = 1 + \frac{1}{r_3}$. The control parameters will be defined after Eqs. (8) and (10).

$$s_{ih} = \dot{e}_{h_{s_i}} + c_{1ih}sig^{n_{1ih}}(e_{h_{s_i}}) + c_{2ih}sig^{n_{2ih}}(e_{h_{s_i}}) \tag{8}$$

where $i = (1, 2, 3)$ are the number of the error model system, and $h = (1, 2, 3)$ are of subsystem numbers of each error model system, and c_{1ih}, c_{2ih} are constant and positive parameters, and n_{1ih}, n_{2ih} are constants positive and smaller than one obtained from the expressed equations in (9).

$$\begin{cases} n_{i1} = N \\ n_{i2} = \frac{N}{2-N} \end{cases}; N = (0, 1) \tag{9}$$

and

$$\begin{cases} u_{ih} = u_{eqih} + u_{r_{ih}} \\ u_{eqih} = f_{h_m} - f_{h_{s_i}} - c_{1ih}sig^{n_{1ih}}(e_{h_{s_i}}) - c_{2ih}sig^{n_{2ih}}(e_{h_{s_i}}) \\ \dot{u}_{r_{ih}} = -k_{1hi}sig^{\alpha_{1hi}}(s_{ih}) - k_{2hi}sig^{\alpha_{2hi}}(s_{ih}) \end{cases} \tag{10}$$

In these inputs k_{1hi}, k_{2hi} are positive constants and $\alpha_{1hi}, \alpha_{2hi}$ are positive constants and smaller than one.

Proof: To prove the finite time stability of designed control inputs by the NTSMC method, we must first prove the finite time stability of the sliding surfaces, then prove of reaching the system to the sliding surfaces in a finite time by applying designed control inputs to the system.

The finite time stability of the presented sliding surfaces in (8) has been proven in [7, 22], and the only condition of their globally finite time stability is choosing c_{1ih}, c_{2ih} parameters properly in such a way that the polynomial $p^2 + c_{2ih}p + c_{1ih} = 0$ is Horowitz.

To proof of reaching the system to the sliding surfaces in a finite time, we assume the Lyapunov candidate function as $V(x) = \frac{1}{2} \sum_{h=1}^3 s_{ih}$. This Lyapunov candidate function has the conditions of Lemma 1 for the Lyapunov function. We take the derivative of this Lyapunov function and then substitute the value of \dot{e}_{hs_i} from Eq. (7), and applying the control input into it, there comes

$$\dot{V}(x) \leq \sum_{i=1}^3 s_{ih}(-k_{1hi} \text{sig}^{\alpha_{1hi}}(s_{ih}) - k_{2hi} \text{sig}^{\alpha_{2hi}}(s_{ih})) \quad (11)$$

After simplifying the inequality, we have

$$\dot{V}(x) \leq \sum_{i=1}^3 -k_{1hi} |s_{ih}|^{\alpha_{1hi}+1} - k_{2hi} |s_{ih}|^{\alpha_{2hi}+1} \quad (12)$$

In the end, according to Lemma 2 and choosing the values:

$r_1 = (\sqrt{2})^{\alpha_{1hi}+1} k_{1hi} > 0, r_2 = (\sqrt{2})^{\alpha_{2hi}+1} k_{2hi} > 0, r_4 = \alpha_{1hi} = 1 - \frac{1}{r_3}, r_5 = \alpha_{1hi} = 1 + \frac{1}{r_3}$, we rewrite Eq. (12) as follows:

$$\dot{V}(x) \leq -r_1 V^{r_4} - r_2 V^{r_5} \quad (13)$$

And according to Lemma 1, the system is globally finite time stable, and its settling time is obtained from the following inequality $T \leq \pi r_3 (\sqrt{r_1 r_2})^{-1}$. ■

Remark 1: in many real systems, the parameters are not precisely calculable, and there are uncertainties in them, and the expressed systems in this article can also have these parameter uncertainties. Assuming that the uncertainties of the system are modeled as $d_{hi}(t, x)$, consequently, the error dynamics can be rewritten as follows:

$$\dot{e}_{hs_i} = f_{hs_i} - f_{hm} + u_{hs_i} + d_{hi}(t, x) \quad (14)$$

Although there is no exact value for uncertainties, we can obtain a high boundary in the case of uncertainties. Here also assumed that the upper bound of the uncertainties model and its derivative is available, and they are considered as $\|d_{hi}(t, x)\| \leq \eta_{1ih}$ and

$\|\dot{d}_{hi}(t, x)\| \leq \eta_{2ih}$. To design the controller for this new model, it is sufficient to rewrite $\dot{u}_{r_{ih}}$, as follows:

$$\dot{u}_{r_{ih}} = -k_{1hi} \text{sig}^{\alpha_{1hi}}(s_{ih}) - k_{2hi} \text{sig}^{\alpha_{2hi}}(s_{ih}) - \eta_{2ih} \text{sgn}(s_{ih}) \tag{15}$$

It can be proved that this control input also ensures the design goal, providing finite time stability and chattering-free.

Remark 2: If the design goal is to create an anti-synchronization, it is sufficient to define tracking errors for each system as $e_{1s_i} = x_{s_i} + x_m$, $e_{2s_i} = y_{s_i} + y_m$ and $e_{3s_i} = z_{s_i} + z_m$. At this time, slave systems do not behave like the master system behavior but behaved quite symmetrically with it. To fulfill this goal, it can be proven that control inputs are designed as follows:

$$\begin{cases} u_{ih} = u_{eq_{ih}} + u_{r_{ih}} \\ u_{eq_{ih}} = -f_{h_m} - f_{h_{s_i}} - c_{1ih} \text{sig}^{n_{1ih}}(e_{h_{s_i}}) - c_{2ih} \text{sig}^{n_{2ih}}(e_{h_{s_i}}) \\ \dot{u}_{r_{ih}} = -k_{1hi} \text{sig}^{\alpha_{1hi}}(s_{ih}) - k_{2hi} \text{sig}^{\alpha_{2hi}}(s_{ih}) \end{cases} \tag{16}$$

5 Numerical Simulation

The accuracy of designs and proofs can be seen by choosing the given control coefficients in (17) and simulation of the systems and control inputs in MATLAB software. The ode4 with a sample time of 0.001 is the solver of the simulations.

$$\begin{cases} c_{1ih} = 7, c_{2ih} = 10, k_{1ih} = k_{2ih} = 0.5 \\ n_{1ih} = \alpha_{1ih} = 0.4, n_{2ih} = \alpha_{2ih} = 0.2 \end{cases} \tag{17}$$

In Fig. 1, the master system with its slave systems is displayed in three dimensions, and in Figs. 2, 3 and 4, the corresponding state variables of the systems are plotted.

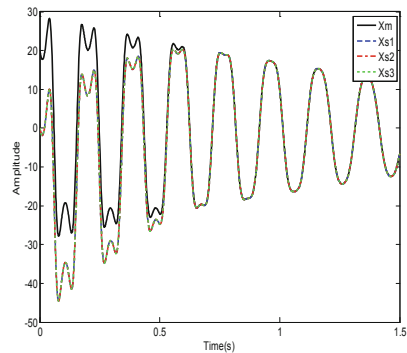
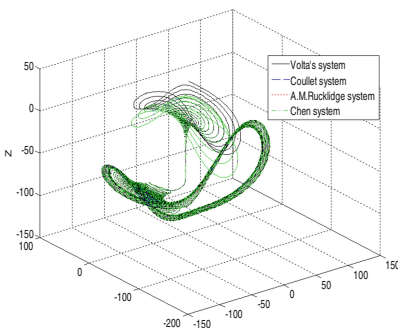


Fig. 1. Master system and slave systems in three dimensions

Fig. 2. The x variable of the systems

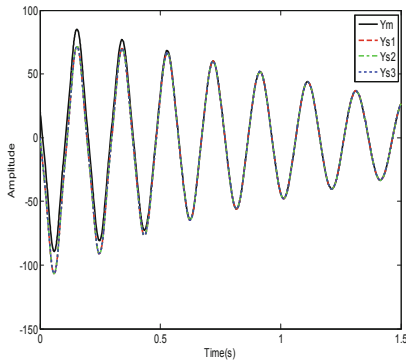


Fig. 3. The y variable of the systems

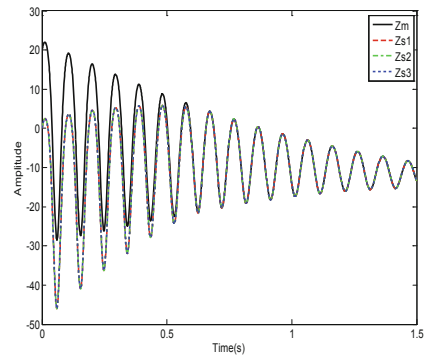


Fig. 4. The z variable of the systems

6 Conclusion

In this paper, four different communication systems have been synchronized with each other by the NTSMC method. The main objective is to ensure secure and reliable communications in this note. The designs aim towards providing globally finite time stability of the systems, which leads to speed up tracking the slave systems of the master system. An extremely important feature of the designs of this article is the removal of undesirable chattering problems, which is essential in the application of secure communication. It is shown that if parameter uncertainties exist in the system, it is possible to upgrade the controllers to robust controllers. It is also available in a non-existent system; it would also enhance controllers to robust controllers. Each designed control input has a large number of control parameters that can control the stability time (settling time) by proper change and consequently control the speed of reaching stability according to the control goal.

Acknowledgments. Andrew Ordys acknowledges support from the National Agency of Academic Exchange (NAWA), “Polish Returns,” grant No: PPN/PPO/2018/1/00063/U/00001.

Ali Soltani Sharif Abadi acknowledges support from Warsaw University of Technology (WUT), grant No: 504440200003.

References

1. Dragomir, D., Gheorghe, L., Costea, S., Radovici, A.: A survey on secure communication protocols for IoT systems. In: 2016 International Workshop on Secure Internet of Things (SIoT), pp. 47–62. IEEE (2016)
2. Sklyarov, V., Skliarova, I.: Secure design of communication networks. In: 2017 2nd International Conference on Anti-Cyber Crimes (ICACC), pp. 193–198. IEEE (2017)
3. Fowler, T.B.: Application of stochastic control techniques to chaotic nonlinear systems. *IEEE Trans. Autom. Control* **34**(2), 201–205 (1989)
4. Bing, Q., Liang-Rui, T., Jing, L., Yi, S.: A new chaotic secure communication system (2008)

5. Li, Z., Li, K., Wen, C., Soh, Y.C.: A new chaotic secure communication system. *IEEE Trans. Commun.* **51**(8), 1306–1312 (2003)
6. Hosseinabadi, P.A., Abadi, A.S.S., Mekhilef, S.: Fuzzy adaptive finite-time sliding mode controller for trajectory tracking of ship course systems with mismatched uncertainties. *Int. J. Autom. Control* **16**(3–4), 255–271 (2022)
7. Feng, Y., Han, F., Yu, X.: Chattering free full-order sliding-mode control. *Automatica* **50**(4), 1310–1314 (2014)
8. Abadi, A.S.S., Hosseinabadi, P.A., Mekhilef, S.: Fuzzy adaptive fixed-time sliding mode control with state observer for a class of high-order mismatched uncertain systems. *Int. J. Control Autom. Syst.* **18**, 2492–2508 (2020)
9. Dasgupta, T., Paral, P., Bhattacharya, S.: Fractional order sliding mode control based chaos synchronization and secure communication. In: 2015 International Conference on Computer Communication and Informatics (ICCCI), pp. 1–6. IEEE (2015)
10. Rodrigues, V.H.P., Oliveira, T.R.: Chaos synchronization applied to secure communication via sliding mode control and norm state observers. In: 2014 13th International Workshop on Variable Structure Systems (VSS), pp. 1–6. IEEE (2014)
11. Shen, L.-Q., Ma, J.-W., Liu, L., Du, H.-Y., Zhang, P.: Adaptive sliding mode synchronization of a class of chaotic systems and its application in secure communication. In: 2013 32nd Chinese Control Conference (CCC), pp. 956–961. IEEE (2013)
12. Zhang, J., Hu, Q., Jiang, C., Ma, G.: Robust sliding mode attitude synchronization control of satellite formation with communication time-delay. In: 2013 32nd Chinese Control Conference (CCC), pp. 7005–7010. IEEE (2013)
13. Alinaghi Hosseinabadi, P., Soltani Sharif Abadi, A., Schwartz, H., Pota, H., Mekhilef, S.: Fixed-time sliding mode observer-based controller for a class of uncertain nonlinear double integrator systems. *Asian J. Control* **25**, 3052 (2023)
14. Hosseinabadi, P.A., Pota, H., Mekhilef, S., Schwartz, H.: Fixed-time observer-based control of DFIG-based wind energy conversion systems for maximum power extraction. *Int. J. Electr. Power Energy Syst.* **146**, 108741 (2023)
15. Abadi, A.S.S., Mehrizi, M.H., Hosseinabadi, P.A.: Fuzzy adaptive terminal sliding mode control of SIMO nonlinear systems with TS fuzzy model. In: 2018 6th Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS), pp. 185–189. IEEE (2018)
16. Parsegov, S., Polyakov, A., Shcherbakov, P.: Fixed-time consensus algorithm for multi-agent systems with integrator dynamics. *IFAC Proceed.* Vol. **46**(27), 110–115 (2013)
17. Alinaghi Hosseinabadi, P., Soltani Sharif Abadi, A., Mekhilef, S., Pota, H.R.: Fixed-time adaptive robust synchronization with a state observer of chaotic support structures for offshore wind turbines. *J. Control Autom. Electr. Syst.* **32**(4), 942–955 (2021)
18. Abadi, A.S.S., Hosseinabadi, P.A., Mekhilef, S.: Two novel AOTSMC of photovoltaic system using VSC model in smart grid. In: 2017 Smart Grid Conference (SGC), pp. 1–6. IEEE (2017)
19. Hu, J.-B., Han, Y., Zhao, L.-D.: Synchronization in the Genesio Tesi and Couillet systems using the backstepping approach. In: *J. Phys. Conf. Ser.* 2008 **96**(1), 012150 (2008). IOP Publishing (2008)
20. Ramanathan, C., et al.: A new chaotic attractor from Rucklidge system and its application in secured communication using OFDM. In: 2017 11th International Conference on Intelligent Systems and Control (ISCO), pp. 241–245. IEEE (2017)
21. Liang, X., Qi, G.: Mechanical analysis of Chen chaotic system. *Chaos Solitons Fract.* **98**, 173–177 (2017)
22. Bhat, S.P., Bernstein, D.S.: Geometric homogeneity with applications to finite-time stability. *Math. Control Sig. Syst. (MCSS)* **17**(2), 101–127 (2005)



Design of Robust H_∞ Control of an Active Inerter-Based Vehicle Suspension System

Keyvan Karim Afshar^(✉), Roman Korzeniowski, and Jarosław Konieczny

Department of Process Control, AGH University of Science and Technology, Krakow, Poland
afshar@agh.edu.pl

Abstract. In this study, a robust H_∞ controller for a quarter-car model of an active inerter-based suspension in the presence of external disturbance has been investigated. Its prime goal is to improve the inherent trade-offs among ride quality, handling performance, suspension travel, and power consumption. Inerters have been widely used to suppress undesirable vibrations of various types of mechanical structures. First, the dynamics and state space of the active inerter-based suspension system were achieved for the quarter-car model. To meet the specified objectives, and guarantee the prescribed disturbance attenuation level of the closed-loop system, the Lyapunov stability function, and linear matrix inequality (LMI) approaches have been used to fulfill the robust H_∞ criterion. Furthermore, to limit the gain of the controller, some LMIs have been added. Numerical simulations show that an active suspension based on inerter performs much better than a passive suspension with inerter and active suspension without inerter.

Keywords: Active inerter-based suspension system · Quarter-car model · Robust H_∞ control, Linear matrix inequality

1 Introduction

The primary goal in the development of vehicle suspension systems, is to reduce acceleration of the car body and passengers while ensuring good contact between tires and the road. The suspension travel must also be limited in the permissible working space. These objectives - ride comfort, road holding, and suspension travel - can be in opposition with each other, and the design problem is to find a compromise between them [1]. In terms of the control structure for suspension systems, there are three main categories which have been developed to attain the required performance of the vehicle: passive, semi-active and active suspension systems [1, 2]. Many investigations have indicated that the active suspension system is an effective method to improve suspension performance, although it demands external energy [3].

To avoid the consequences of vibration, many techniques have been introduced, such as isolating systems from vibration, controlling systems, redesigning systems to change their natural frequencies, using tuned mass dampers or absorbers, and more [4, 5]. Tuned mass dampers (TMDs) are widely used to suppress unwanted vibrations of various mechanical structures, e.g., buildings, bridges, motorcycles steering system,

vehicle and train suspensions, etc. [4]. The classical TMD consists of mass on a linear spring, and it is known that the classical TMD is especially effective in reducing the response of the main structure in principal resonance, but at other frequencies (even near the resonance frequency) it increases the amplitude of the system's motion [4,5]. This problem is capable to be minimized by novel TMDs containing inerter or magnetorheological dampers, which are intensively developed nowadays. An inerter is a device with two free-moving terminals whose generated force is proportional to the relative acceleration of its terminals. The proportional constant is called the inertance with the unit kilogram [6]. The inerter possesses the effect of mass amplification and would provide much greater inertia compared to its own mass, thus increasing the inertia of the entire dynamic system rather than increasing the mass [5,7]. Due to its mechanical properties, it is therefore an efficient structure for damping vibrations.

The rack-and-pinion, ball-screw, and hydraulic (or fluid) inerters are the three most commonly used inerters. When the inertance is fixed, the inerter is passive; when the inertance can be adjusted, the inerter is semi-active [4]. In [6], both semi-active damper and semi-active inerter were applied in the vehicle suspension system, and a feedback control approach was proposed. In [2], the active inertia-based suspension has used a controllable actuator to produce the desired force. Compared to passive and semi-active suspensions, it provides better dynamic performance, although it requires the most energy. It is noteworthy that in [8] an inerter-based electromagnetic device was presented and implemented in the vehicle suspension system. The proposed device not only improves the performance of the suspension but also generates an amount of electrical energy that can be utilized by other parts of the vehicle, especially the energy required to operate the actuator.

On the other hand, to find a compromise between the conflicting performances, many approaches to active suspension control have been proposed based on various control techniques, such as fuzzy logic and neural network control [9,10], adaptive control [1,10], model predictive control [11], sliding mode control [12], H_∞ control [1,3,11], etc. In particular, the application of H_∞ control of the active vehicle suspension system has been intensively investigated in the context of robustness and damping of road disturbances.

In this paper, the active inerter-based quarter-car suspension system is investigated based on the parallel-connected configuration, since this configuration is simple and space-saving [7]. The H_∞ control (energy-to-energy) is used to optimize the performance requirements of the suspension system. Sufficient stability conditions and performance criteria are derived in the form of LMIs employing the direct Lyapunov method. Moreover, additional LMIs are added to the original condition to reduce the gain of the controller, which results in avoiding measurement noise amplification and saturation of the actuator. The stability conditions are derived as linear matrix inequalities (LMIs) and therefore the stabilization gain of the system is obtained by solving the convex optimization problem.

The subsequent parts of this paper are structured into four sections. The description of the active inerter-based quarter-car suspension system is presented in Sect. 2. Section 3 contains the problem formulation for robust H_∞ control based on the solvability of LMIs. In Sect. 4, the proposed controller is applied to the inerter-based quarter-

car model for performance evaluation. Finally, the conclusion of our findings is given in Sect. 5.

2 Problem Formulation

The inerter is connected in parallel to the spring and damper between the wheel and chassis. The active inerter-based suspension system, as shown in Fig. 1, can be reduced to a 2DOF system considering the vertical dynamics. The quarter-car model is assembled by one sprung mass (car body) that is connected to an unsprung mass. The unsprung mass is free to move vertically and is confronted with the road disturbance input.

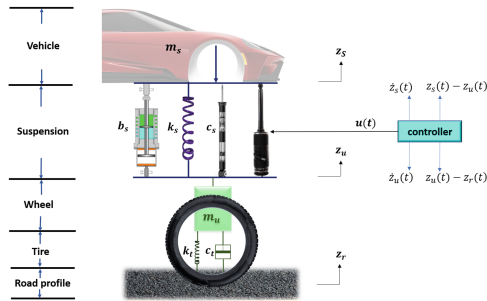


Fig. 1. Active inerter-based quarter-car suspension system.

In Fig. 1, m_s is the mass of the car body, and m_u is the unsprung mass. b_s denotes the inertance of the inerter, c_s represents the damping coefficient of suspension element, and k_s represents the stiffness of the suspension. Likewise, c_t represents the damping coefficient of tire, k_t is the tire stiffness; $u(t)$ denotes actuator force input. $z_s(t)$ represents the vertical displacements of the body, $z_u(t)$ denotes the vertical displacements of the unsprung mass, and $z_r(t)$ denotes the road disturbance input. It is assumed that the tire is in contact with the road at all times, and the characteristics of the suspension elements are linear. The differential equations of motion can be calculated by using Newton's second law as follows

$$m_s \ddot{z}_s(t) + b_s [\ddot{z}_s(t) - \ddot{z}_u(t)] + c_s [\dot{z}_s(t) - \dot{z}_u(t)] + k_s [z_s(t) - z_u(t)] = u(t) \quad (1)$$

$$m_u \ddot{z}_u(t) + b_s [\ddot{z}_u(t) - \ddot{z}_s(t)] + c_s [\dot{z}_u(t) - \dot{z}_s(t)] + k_s [z_u(t) - z_s(t)] + c_t [\dot{z}_u(t) - \dot{z}_r(t)] + k_t [z_u(t) - z_r(t)] = -u(t) \quad (2)$$

Defining four state variables as follow

$$x_1(t) = z_s(t) - z_u(t) \quad , \quad x_2(t) = z_u(t) - z_r(t) \quad , \quad x_3(t) = \dot{z}_s(t) \quad , \quad x_4(t) = \dot{z}_u(t) \quad (3)$$

where $x_1(t)$ is the suspension deflection, $x_2(t)$ is the tire deflection, $x_3(t)$ denotes the vertical velocity of the car body, $x_4(t)$ denotes the vertical velocity of the wheel. Accordingly, by defining $\mathbf{x}(t) = [x_1(t) \ x_2(t) \ x_3(t) \ x_4(t)]^T$, the active inerter-based suspension system can be represented by the following state-space equation

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t) + \mathbf{D}\mathbf{v}(t) \tag{4}$$

where

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 1 \\ -m_u k_s / f & -b_s k_t / f & -m_u c_s / f & (m_u c_s - b_s c_t) / f \\ m_s k_s / f & -(m_s + b_s) k_t / f & m_s c_s / f & -(m_s c_s + (m_s + b_s) c_t) / f \end{bmatrix}$$

$$\mathbf{B} = [0 \ 0 \ m_u / f \ -m_s / f]^T \quad , \quad \mathbf{D} = [0 \ -1 \ b_s c_t / f \ (m_s + b_s) c_t / f]^T$$

$$f = m_s m_u + (m_s + m_u) b_s \quad , \quad \mathbf{v}(t) = \dot{z}_r(t)$$

As mentioned earlier, ride comfort, suspension deflection, and road-holding ability are the three most important performance criteria to consider when developing controllers for vehicle suspension systems. Therefore, the controlled output of the vehicle suspension system can be described by the following state space equation:

$$\mathbf{z}(t) = \mathbf{C}_1 \mathbf{x}(t) + \mathbf{D}_{12} \mathbf{u}(t) + \mathbf{F} \mathbf{v}(t) \tag{5}$$

where

$$\mathbf{C}_1 = \begin{bmatrix} -\rho(m_u k_s / f) & -\rho(b_s k_t / f) & -\rho(m_u c_s / f) & \rho((m_u c_s - b_s c_t) / f) \\ \alpha & 0 & 0 & 0 \\ 0 & \beta & 0 & 0 \end{bmatrix}$$

$$\mathbf{D}_{12} = [\rho(m_u / f) \ 0 \ 0]^T \quad , \quad \mathbf{F} = [\rho(b_s c_t / f) \ 0 \ 0]^T$$

where $\rho > 0$ is a scalar weighting for the ride comfort, $\alpha > 0$ is a scalar weighting for the suspension deflection, and $\beta > 0$ is a scalar weighting for the tyre deflection.

As aforementioned, we assume the case that all the state variables $\mathbf{x}(t)$ can be measured employing appropriate sensors, leading to the design of a state-feedback H_∞ controller.

$$\mathbf{y}(t) = \mathbf{C}_2 \mathbf{x}(t) \quad , \quad \mathbf{C}_2 = \mathbf{I} \tag{6}$$

For the design of an H_∞ controller, Consider the following state-feedback controller

$$\mathbf{u}(t) = \mathbf{K}\mathbf{y}(t) = \mathbf{K}\mathbf{x}(t) \tag{7}$$

where \mathbf{K} is the state feedback gain matrix to be designed, such that, first, the closed-loop system without external disturbance is asymptotically stable, and second, under zero initial condition the \mathcal{L}_2 gain (i.e., H_∞ norm) of the closed-loop system guarantees $\|\mathbf{z}(t)\|_{\mathcal{L}_2}^2 < \gamma^2 \|\mathbf{v}(t)\|_{\mathcal{L}_2}^2$ for all nonzero $\mathbf{v}(t) \in \mathcal{L}_2 [0 \ \infty)$, and some scalar $\gamma > 0$.

3 H_∞ Controller Design

The active inerter-based vehicle suspension system can be defined by the following state-space equations

$$\begin{aligned}\dot{\mathbf{x}}(t) &= \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t) + \mathbf{D}\mathbf{v}(t) \\ \mathbf{z}(t) &= \mathbf{C}_1\mathbf{x}(t) + \mathbf{D}_{12}\mathbf{u}(t) + \mathbf{F}\mathbf{v}(t) \\ \mathbf{y}(t) &= \mathbf{C}_2\mathbf{x}(t)\end{aligned}\quad (8)$$

where $\mathbf{x}(t) \in R^n$ is the state, $\mathbf{u}(t) \in R^m$ is the input vector, $\mathbf{y}(t) \in R^p$ is the measured output, $\mathbf{z}(t) \in R^d$ is the controlled output, $\mathbf{v}(t) \in R^q$ is the external disturbance vector, matrices \mathbf{A} , \mathbf{B} , \mathbf{D} , \mathbf{C}_1 , \mathbf{D}_{12} , \mathbf{F} , and \mathbf{C}_2 , are all constant real matrices with appropriate dimensions. In this section, we will solve the problem of state-feedback robust H_∞ controller for active inerter-based suspension system.

Assumption 1. *In this paper, it is assumed that the external disturbance signal $\mathbf{v}(t)$ is square-integrable, that is*

$$\|\mathbf{v}(t)\|_{\mathcal{L}_2}^2 = \int_0^\infty \|\mathbf{v}(s)\|^2 ds < v_{max} < \infty$$

Lemma 1. [1] *Given constant matrices Ω_1 , Ω_2 and Ω_3 satisfying $\Omega_1 = \Omega_1^T$ and $\Omega_2 > 0$, then $\Omega_1 + \Omega_3^T \Omega_2^{-1} \Omega_3 < 0$, if and only if*

$$\begin{bmatrix} \Omega_1 & \Omega_3^T \\ \Omega_3 & -\Omega_2 \end{bmatrix} < 0$$

Theorem 1. *Assuming positive constants γ , L_R , and L_S , the linear suspension system (Eq. (8)) with state-feedback controller (Eq. (7)) in the absence of external disturbance is asymptotically stable and in the presence of external disturbance satisfies $\|\mathbf{z}(t)\|_{\mathcal{L}_2}^2 < \gamma^2 \|\mathbf{v}(t)\|_{\mathcal{L}_2}^2$ for $\mathbf{v}(t) \in \mathcal{L}_2 [0 \infty)$, if there exist symmetric positive definite matrix $\mathbf{X} > 0$ and matrix \mathbf{Y} with appropriate dimensions, such that the following LMIs hold*

$$\begin{bmatrix} \mathbf{A}\mathbf{X} + \mathbf{X}\mathbf{A}^T + \mathbf{Y}^T\mathbf{B}^T + \mathbf{B}\mathbf{Y} & \mathbf{X}\mathbf{C}_1^T\mathbf{F} + \mathbf{Y}^T\mathbf{D}_{12}^T\mathbf{F} + \mathbf{D} & \mathbf{X}\mathbf{C}_1^T + \mathbf{Y}^T\mathbf{D}_{12}^T \\ \mathbf{F}^T\mathbf{C}_1\mathbf{X} + \mathbf{F}^T\mathbf{D}_{12}\mathbf{Y} + \mathbf{D}^T & \mathbf{F}^T\mathbf{F} - \gamma^2\mathbf{I} & \mathbf{0} \\ \mathbf{C}_1\mathbf{X} + \mathbf{D}_{12}\mathbf{Y} & \mathbf{0} & -\mathbf{I} \end{bmatrix} < 0 \quad (9)$$

$$\begin{bmatrix} L_R\mathbf{I} & \mathbf{Y}^T \\ \mathbf{Y} & \mathbf{I} \end{bmatrix} > 0 \quad (10)$$

$$\begin{bmatrix} L_S\mathbf{I} & \mathbf{I} \\ \mathbf{I} & \mathbf{X} \end{bmatrix} > 0 \quad (11)$$

In this case, if inequalities (9)–(11) have a feasible solution, the stabilizing gain of the state-feedback controller (Eq. (7)) is given by $\mathbf{K} = \mathbf{Y}\mathbf{X}^{-1}$.

Proof. Choose the following Lyapunov function

$$V(t) = \mathbf{x}^T(t) \mathbf{p} \mathbf{x}(t) > 0 \tag{12}$$

and $\mathbf{p} = \mathbf{p}^T > 0$ is the matrix to be chosen. The derivative of $V(t)$ is taken as

$$\begin{aligned} \dot{V}(t) &= \dot{\mathbf{x}}^T(t) \mathbf{p} \mathbf{x}(t) + \mathbf{x}^T(t) \mathbf{p} \dot{\mathbf{x}}(t) = \mathbf{x}^T(t) (\mathbf{A} + \mathbf{BK})^T \mathbf{p} \mathbf{x}(t) + \mathbf{v}^T(t) \mathbf{D}^T \mathbf{p} \mathbf{x}(t) \\ &\quad + \mathbf{x}^T(t) \mathbf{p} (\mathbf{A} + \mathbf{BK}) \mathbf{x}(t) + \mathbf{x}^T(t) \mathbf{p} \mathbf{D} \mathbf{v}(t) < 0 \end{aligned} \tag{13}$$

Assuming zero initial condition ($\mathbf{u}(t) = 0$), we have $V(t)|_{t=0} = 0$. Now, consider the following index

$$J_\infty = \int_0^\infty \left[\mathbf{z}(t)^T \mathbf{z}(t) - \gamma^2 \mathbf{v}(t)^T \mathbf{v}(t) \right] dt \tag{14}$$

Then, for any nonzero $\mathbf{v}(t) \in \mathcal{L}_2 [0 \infty)$, there holds

$$\begin{aligned} J_\infty &\leq \int_0^\infty \left[\mathbf{z}(t)^T \mathbf{z}(t) - \gamma^2 \mathbf{v}(t)^T \mathbf{v}(t) \right] dt + V(t)|_{t=\infty} - V(t)|_{t=0} \\ &= \int_0^\infty \left[\mathbf{z}(t)^T \mathbf{z}(t) - \gamma^2 \mathbf{v}(t)^T \mathbf{v}(t) + \dot{V}(t) \right] dt = \int_0^\infty \zeta^T \Pi_1 \zeta dt \end{aligned} \tag{15}$$

where $\zeta = [\mathbf{x}(t)^T \ \mathbf{v}(t)^T]^T$, and

$$\Pi_1 = \begin{bmatrix} (\mathbf{A} + \mathbf{BK})^T \mathbf{p} + \mathbf{p} (\mathbf{A} + \mathbf{BK}) + (\mathbf{C}_1 + \mathbf{D}_{12} \mathbf{K})^T (\mathbf{C}_1 + \mathbf{D}_{12} \mathbf{K}) & (\mathbf{C}_1 + \mathbf{D}_{12} \mathbf{K})^T \mathbf{F} + \mathbf{p} \mathbf{D}^T \\ \mathbf{F}^T (\mathbf{C}_1 + \mathbf{D}_{12} \mathbf{K}) + \mathbf{D} \mathbf{p} & \mathbf{F}^T \mathbf{F} - \gamma^2 \mathbf{I} \end{bmatrix} \tag{16}$$

Assuming the zero-disturbance input, that is, $\mathbf{v}(t) = 0$; if Eq. (16) is negative-definite, that is, $\Pi_1 < 0$, then $\dot{V}(t) < 0$ and the asymptotic stability of system Eq. (8) is guaranteed. When $\mathbf{v}(t) \in \mathcal{L}_2 [0 \infty)$, and $\Pi_1 < 0$, this implies that $J_\infty < 0$, and therefore $\|\mathbf{z}(t)\|_{\mathcal{L}_2}^2 < \gamma^2 \|\mathbf{v}(t)\|_{\mathcal{L}_2}^2$.

By utilizing Lemma 1 (Schur complement), $\Pi_1 < 0$ is equivalent to

$$\Pi_2 = \begin{bmatrix} (\mathbf{A} + \mathbf{BK})^T \mathbf{p} + \mathbf{p} (\mathbf{A} + \mathbf{BK}) & (\mathbf{C}_1 + \mathbf{D}_{12} \mathbf{K})^T \mathbf{F} + \mathbf{p} \mathbf{D} & (\mathbf{C}_1 + \mathbf{D}_{12} \mathbf{K})^T \\ \mathbf{F}^T (\mathbf{C}_1 + \mathbf{D}_{12} \mathbf{K}) + \mathbf{D} \mathbf{p} & \mathbf{F}^T \mathbf{F} - \gamma^2 \mathbf{I} & \mathbf{0} \\ (\mathbf{C}_1 + \mathbf{D}_{12} \mathbf{K}) & \mathbf{0} & -\mathbf{I} \end{bmatrix} < 0 \tag{17}$$

Pre- and post-multiplying Eq. (17) by $diag(\mathbf{p}^{-1}, \mathbf{I}, \mathbf{I})$ and its transpose, respectively, we obtain

$$\Pi_3 = \begin{bmatrix} \mathbf{A} \mathbf{p}^{-1} + \mathbf{p}^{-1} \mathbf{A}^T + \mathbf{p}^{-1} \mathbf{K}^T \mathbf{B}^T + \mathbf{BK} \mathbf{p}^{-1} & \mathbf{p}^{-1} \mathbf{C}_1^T \mathbf{F} + \mathbf{p}^{-1} \mathbf{K}^T \mathbf{D}_{12}^T \mathbf{F} + \mathbf{D} & \mathbf{p}^{-1} \mathbf{C}_1^T + \mathbf{p}^{-1} \mathbf{K}^T \mathbf{D}_{12}^T \\ \mathbf{F}^T \mathbf{C}_1 \mathbf{p}^{-1} + \mathbf{F}^T \mathbf{D}_{12} \mathbf{K} \mathbf{p}^{-1} + \mathbf{D}^T & \mathbf{F}^T \mathbf{F} - \gamma^2 \mathbf{I} & \mathbf{0} \\ \mathbf{C}_1 \mathbf{p}^{-1} + \mathbf{D}_{12} \mathbf{K} \mathbf{p}^{-1} & \mathbf{0} & -\mathbf{I} \end{bmatrix} < 0 \tag{18}$$

After substituting $\mathbf{X} = \mathbf{p}^{-1}$, $\mathbf{Y} = \mathbf{K} \mathbf{p}^{-1}$ into Eq. (18) we obtain Eq. (9). Conditions $\mathbf{X} > 0$ and Eq. (9) guarantee $\Pi_1 < 0$, which further implies that $J_\infty < 0$ in Eq. (15), and therefore $\|\mathbf{z}(t)\|_{\mathcal{L}_2}^2 < \gamma^2 \|\mathbf{v}(t)\|_{\mathcal{L}_2}^2$.

In the real implementation of control systems (including active suspension systems), direct consequences of high gain control can lead to some major problems such as actuator saturation and noise amplification. Therefore, the gain matrix \mathbf{K} should be limited. In this study, we follow the methodology utilized for this problem in [3]. Accordingly, conforming to expression $\mathbf{K} = \mathbf{Y}\mathbf{X}^{-1}$, we can restrict the size of the gain matrix \mathbf{K} by constraining the two matrices \mathbf{Y} and \mathbf{X}^{-1} . We assigned

$$\mathbf{Y}^T \mathbf{Y} < L_R \mathbf{I}, \quad L_R > 0 \quad (19)$$

$$\mathbf{X}^{-1} < L_S \mathbf{I}, \quad L_S > 0 \quad (20)$$

Utilizing Lemma 1, LMIs (19) and (20) lead to the LMIs (10) and (11), respectively. Therefore, the proof of Theorem 1 is completed.

4 Application to Inerter-Based Quarter-Car Suspension Control

In this section, we will apply the proposed method to the inerter-based quarter-car model described in Sect. 2 to illustrate the effectiveness of the proposed method. The parameters of the inerter-based quarter-car model are listed in Table 1.

Table 1. System parameter values of the inerter-based quarter-car suspension model.

Parameter	Value	Parameter	Value
m_s	972.2 kg	m_{it}	113.6 kg
k_s	42,719.6 N/m	c_s	1095 Ns/m
k_t	101,115 N/m	c_t	14.6 Ns/m
b_s	6 kg		

By setting $\gamma = 7.5$, $L_R = 8 \times 10^5$, $L_S = 8 \times 10^2$, $\alpha = 32$, $\beta = 8$, $\rho = 0.8$ and solving the convex optimization problem formulated in Theorem 1 using the YALMIP toolbox [13], the gain matrix of controller is obtained

$$\mathbf{K}_I = [-44061.95 \quad -24605.65 \quad -12152.41 \quad -370.38]$$

And for brevity, we will indicate the proposed controller as Controller **I** hereafter. To assess the performance of the proposed controller, acquired results will be compared to the results obtained by robust H_∞ control for active vehicle suspension without inerter, and will be denoted as Controller **II** for brevity. The obtained controller gain with design parameter $\gamma = 7.5$ is as follows

$$\mathbf{K}_{II} = [-170771.42 \quad -128699.69 \quad -41770.58 \quad -1395.68]$$

According to ISO 2361, improving ride comfort is equivalent to minimizing the vertical acceleration of a vehicle system in the frequency range from 4 Hz to 8 Hz [1]. Therefore, we first focus on the frequency responses from ground velocity to vertical body

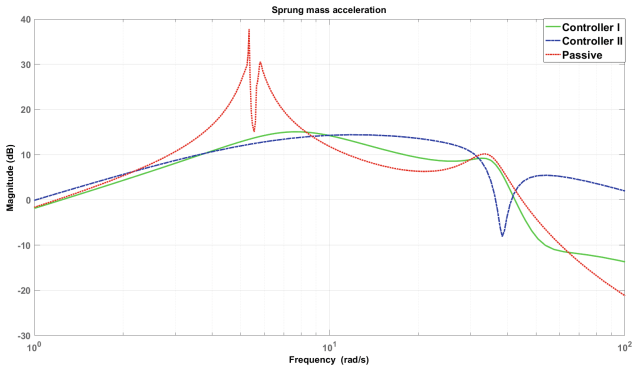


Fig. 2. Frequency responses for the open- and closed-loop systems

acceleration for the passive and closed-loop systems using the robust H_∞ state-feedback controllers. From Fig. 2, we can see that the desired controller I and the controller II can provide the lower value of the H_∞ norm over the frequency range of 4 Hz–8 Hz.

Performance of the quarter-car suspension system is capable to be assessed by examining three response quantities, that is, the sprung mass acceleration $\dot{x}_3(t)$, the suspension deflection $x_1(t)$, and the tire deflection $x_2(t)$. In the following sub-sections, we will utilize Shock (Bump) and Vibration (Rough Road) road profiles to evaluate the performance of the inerter-based quarter-car suspension system.

4.1 Bump Response

Here the bumps or potholes with relatively short duration and high intensity confronted in a smooth surface was considered to reveal the transient response characteristic, which is given by

$$z_{rf}(t) = \begin{cases} \frac{a}{2} \left(1 - \cos\left(\frac{2\pi v_0}{l} t\right) \right) & , 0 \leq t \leq \frac{l}{v_0} \\ 0 & , t > \frac{l}{v_0} \end{cases} \quad (21)$$

where a and l are the height and the length of the bump. We choose $a = 0.1 \text{ m}$, $l = 2 \text{ m}$ and the vehicle forward velocity as $v_0 = 45 \text{ km/h}$.

The response of the quarter-car suspension system with inerter by using Controller I and without inerter by using Controller II, and passive suspension are compared in Fig. 3. It displays the sprung mass acceleration, suspension deflection, tire deflection, and control effort of the active controllers. It can be seen from Fig. 3 that the Controllers I and II compared to the passive suspension acquire better responses. The simulation results confirm that the active suspension with inerter is better than the active suspension without inerter with respect to all performance criteria for the bump disturbance. In addition, the required control effort for Controller I is less than for Controller II.

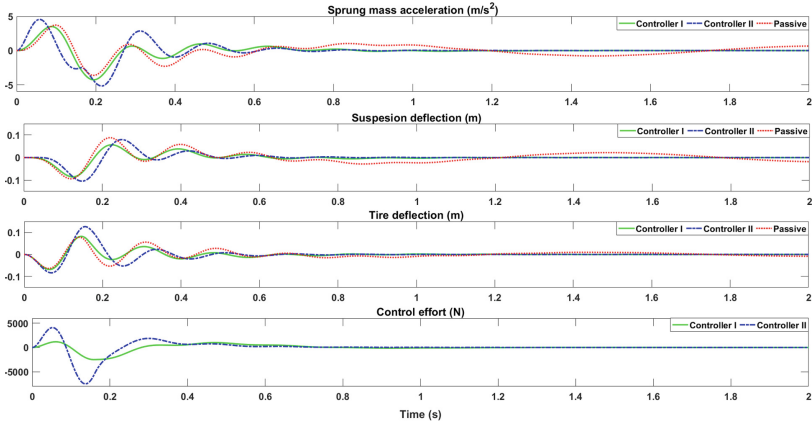


Fig. 3. Sprung mass acceleration, Suspension deflection, Tire deflection, Control effort

To qualitatively evaluate the control effort of two active control methods, their energy consumption is calculated through the following \mathcal{L}_2 norm value:

$$\|\mathbf{u}(t)\|_{\mathcal{L}_2} = \sqrt{\int_0^{\bar{T}} \mathbf{u}(t)^T \mathbf{u}(t) dt} \quad (22)$$

where \bar{T} is the simulation time. Energy consumption of two controllers is shown in Table 2. It can be seen that the active suspension system with inerter has a superior performance compared to the active suspension without inerter and low gain of the Controller I lead to the less required energy consumption.

Table 2. Energy consumption of active controllers ($\bar{T} = 2s$)

	Controller I	Controller II
Energy consumption	891.92	1971.4

4.2 Random Response

Generally, it is capable to assume random vibrations as road disturbances, which are consistent and typically specified as the random process. The ground displacement power spectral density (PSD) is defined as follows [3]

$$S_g(\Omega) = \begin{cases} S_g(\Omega_0) \left(\frac{\Omega}{\Omega_0}\right)^{-n_1} & \text{if } \Omega \leq \Omega_0 \\ S_g(\Omega_0) \left(\frac{\Omega}{\Omega_0}\right)^{-n_2} & \text{if } \Omega > \Omega_0 \end{cases} \quad (23)$$

where $\Omega_0 = 1/2\pi$ stands for reference spatial frequency and Ω is a spatial frequency. The value of $S_g(\Omega_0)$ denotes a measure for the roughness coefficient of the road. n_1 and n_2 represent the road roughness constants.

In particular, if the vehicle is presumed to travel with a constant horizontal speed v_0 over a given road, it is capable to simulate the force resulting from the road irregularities by the following series

$$z_r f(t) = \sum_{n=1}^N s_n \sin(n\omega_0 t + \varphi_n) \tag{24}$$

where $s_n = \sqrt{2s_g(n\Delta\Omega)\Delta\Omega}$, $\Delta\Omega = 2\pi/L$, and L is the length of the road segment considered. The amplitudes s_n of the excitation harmonics are assessed from the road spectra selected. Additionally, the value of the fundamental temporal frequency ω_0 is determined from $\omega_0 = \frac{2\pi}{L}v_0$. While the phases φ_n are treated as random variables, following a uniform distribution in the interval $[0, 2\pi)$.

According to ISO 2631 standards, road class D (poor quality) ($s_g(\Omega_0) = 256 \times 10^{-6} m^3$), is selected as a typical road profile. In this paper, $n_1 = 2, n_2 = 1.5, L = 100, N_f = 200$ and the horizontal speed $v_0 = 36 m/s$, are utilized to generate the random road profile.

The Monte Carlo simulation is utilized to assess the probabilistic characteristics of the random response. Therefore, taking into account the random variable φ_n of the excitation applied, the performance index of the Root Mean Square (RMS) is determined by the expected values:

$$J_1 = E \left[\frac{1}{\bar{T}} \int_0^{\bar{T}} [\dot{x}_3(t)]^2 dt \right] \tag{25}$$

$$J_2 = E \left[\frac{1}{\bar{T}} \int_0^{\bar{T}} [x_1(t)]^2 dt \right] \tag{26}$$

$$J_3 = E \left[\frac{1}{\bar{T}} \int_0^{\bar{T}} [x_2(t)]^2 dt \right] \tag{27}$$

For sprung mass acceleration J_1 , suspension deflection J_2 , tire deflection J_3 ; where $\bar{T} = L/v_0$ is the temporal measurement period. For calculating RMS values, we have considered $\bar{T} = 5$ in Eqs. (25)–(27) and the simulation has been run randomly 100 times.

To validate the effectiveness of controller **I** in dealing with the active suspension system based on inerter, the RMS ratios $J_{Ii}(t)/J_i(t), J_{Pi}(t)/J_i(t), i = 1, 2, 3$, are calculated, where $J_i(t)$ denotes the RMS value of the active suspension with inerter by using Controller **I**, $J_{Ii}(t)$ denotes the RMS value of the active suspension without inerter by using Controller **II**, and $J_{Pi}(t)$ is the RMS value of the passive suspension with inerter.

Table 3 represents the results of RMS ratios for Controller **I** and Controller **II** of the active quarter-car suspension system with inerter and without inerter, respectively, and passive suspension system with inerter for the poor (class D) quality road profile. The control efforts of the active controllers are also shown in Table 3. It can be seen from Table 3 that the RMS ratios of Controllers **I** and **II** are always less than the passive suspension system (the response ratio is more than 1). On the other hand, it can be seen that sprung mass acceleration, suspension deflection, and tire deflection with Controller **I** acquire better response compared to Controller **II**. In addition, the required control effort for Controller **I** is less than for Controller **II**.

Table 3. Energy consumption and RMS values of random road profile

performance criteria	$J_{II}(t)/J_I(t)$	$J_{PI}(t)/J_I(t)$	
sprung mass acceleration	1.5334	2.3928	
suspension deflection	1.2307	3.0286	
tire deflection	2.1475	2.4339	
	Passive	Controller I	Controller II
energy consumption	—	452.79	654.14

5 Conclusions

In this paper, the performance of an active inerter-based quarter-car suspension system has been investigated. A Robust H_∞ controller is developed to optimize the \mathcal{L}_2 norm of the active inerter-based suspension system to enhance the ride comfort performance, suspension deflections and the tire loads. Furthermore, to restrict the gain of the controller, two additional LMIs are added to the obtained sufficient conditions. Finally, to validate the effectiveness of the proposed approach, it is applied to the active inerter-based quarter-car suspension system to minimize the influence of road disturbance on the system performance. It has been observed that for all performance requirements, the active inerter-based suspension system achieves better response compared to both the active suspension without inerter and passive suspension with inerter.

References

1. Liu, H., Gao, H., Li, P.: Handbook of vehicle suspension control systems. Institution of Engineering and Technology (2013)
2. Sun, W., Pan, H., Zhang, Y., Gao, H.: Multi-objective control for uncertain nonlinear active suspension systems. *Mechatronics* **24**(4), 318–327 (2014). *Vibration control systems*
3. Karim Afshar, K., Javadi, A., Jahed-Motlagh, M.R.: Robust H_∞ control of an active suspension system with actuator time delay by predictor feedback. *IET Control Theory Appl.* **12**, 1012–1023 (2018)
4. Chen, M.Z.Q., Hu, Y.: *Inerter and Its Application in Vibration Control Systems*. Springer, Singapore (2019). <https://doi.org/10.1007/978-981-10-7089-1>
5. Brzeski, P., Kapitaniak, T., Perlikowski, P.: Novel type of tuned mass damper with inerter which enables changes of inertance. *J. Sound Vib.* **349**, 56–66 (2015)
6. Chen, M.Z.Q., Hu, Y., Li, C., Chen, G.: Application of semi-active inerter in semi-active suspensions via force tracking. *J. Vibr. Acoust.* **138**(4), 041014 (2016)
7. Wang, Y., Jin, X. Y., Zhang, Y. S., Ding, H., Chen, L. Q.: Dynamic performance and stability analysis of an active inerter-based suspension with time-delayed acceleration feedback control. *Bullet. Polish Acad. Sci.: Techn. Sci.* **70**(2), e140687 (2022)
8. Hu, Y., Du, H., Chen, M.Z.Q.: An inerter-based electromagnetic device and its application in vehicle suspensions. In: 34th Chinese Control Conference (CCC), pp. 2060–2065 (2015)
9. Kumar, P.S., Sivakumar, K., Kanagarajan, R., Kuberan, S.: Adaptive neuro fuzzy inference system control of active suspension system with actuator dynamics. *J. Vibroengineering* **20**, 541–549 (2018)
10. Taghavifar, H., Mardani, A., Hu, C., Qin, Y.: Adaptive robust nonlinear active suspension control using an observer-based modified sliding mode interval type-2 fuzzy neural network. *IEEE Trans. Intell. Vehicles* **5**(1), 53–62 (2020)

11. Bououden, S., Chadli, M., Karimi, H.R.: A robust predictive control design for nonlinear active suspension systems. *Asian J. Control* **18**(1), 122–132 (2016)
12. Konieczny, J., Sibiak, M., Raczka, W.: Active vehicle suspension with anti-roll system based on advanced sliding mode controller. *Energies* **13**(21), 5560 (2020)
13. Lofberg, J.: YALMIP: a toolbox for modeling and optimization in MATLAB. In: 2004 IEEE International Conference on Robotics and Automation (IEEE Cat. No.04CH37508), pp. 284–289 (2004)



Nonlinear Adaptive Control with Invertible Fuzzy Model

Marcin Jastrzębski[✉], Jacek Kabziński, and Rafal Zawisłak

Lodz University of Technology, Stefanowskiego 18, 90-537 Lodz, Poland
{marcin.jastrzebski,jacek.kabzinski,rafal.zawislak}@p.lodz.pl

Abstract. The presented concept of non-linear adaptive control can be used for any system with complex, nonlinear dependencies of control and state variables. It can be applied even when the plant model is partially represented by digital data only. The main idea is to propose a simple fuzzy model to approximate a nonlinear map of control and state, and to develop a fast numerical inversion technique providing on-line control which ensures the desired output of the fuzzy model. Our aim is to obtain a short execution time, to enable applications in DSP-based, fast control systems. Moreover, on-line adaptation of the model parameters should improve the system performance, even if the initial model is inaccurate. The developed fuzzy inversion technique can cooperate with many nonlinear adaptive control approaches: adaptive backstepping, adaptive model following etc.

1 Introduction

Effective non-linear control techniques have been developed for several decades. Their applications are more and more numerous and common. But usually non-linear control is based on some partial assumptions about linearity. A good example of a such situation is the backstepping or adaptive backstepping technique [1]. It was originally developed for systems in the strict-feedback form, when it is assumed that subsequent subsystems are affine in virtual controls and the complete plant is affine with respect to the available control input. Unfortunately, for numerous real plants it is impossible to create a strict-feedback model. A force acting on a levitating metal object in a magnetic levitation system (it is a nonlinear function of the current (control) and position (state variable)), or a mass flow rate in a flow valve for a pneumatic actuator (it is nonlinear function of the coil voltage (control) and differential pressure (state variable)) are good examples of this situation. Therefore, a pure-feedback form model (assuming nonlinear functions of control and state) can be most widely applied for practical applications [2,3]. But, in this case, any iterative design method must struggle to solve implicit equations at each stage of the control design. Moreover, in many cases the nonlinear dependencies between the control and the state are too complex to derive any analytical model and the information is represented as a set of digital data.

In this contribution we describe a concept of a design technique to cope with such complex cases. The main idea is to propose a simple fuzzy model to approximate a non-linear control and state mapping, based on discrete data, and to develop a fast numerical inversion technique to provide on-line control that will deliver the desired model output. Our aim is to obtain a short execution time, to enable applications in DSP-based, fast control systems. Moreover, on-line adaptation of the model parameters should improve the system performance, even if the initial model is not extremely accurate.

The presented concept may be placed somewhere in the range of approximation techniques in nonlinear control. The use of a fuzzy model of Sugeno type subject to adaptation has been shown, among others, in [4, 5]. However, it concerned only that part of the system unrelated to the control. Another approach was presented in [6], while [7] describes an approximate inversion.

The concepts of the selected fuzzy model and the numerical inversion procedure are described in Sect. 2, where some examples are included. An example of adaptive nonlinear control based on the developed fuzzy inversion is presented in Sect. 3. Although the example concerns backstepping control, the developed fuzzy technique can cooperate with many other nonlinear adaptive control approaches.

2 Invertible Fuzzy Model

We consider a non-linear, continuous map $y = f(x_1, x_2)$, $f : A \times B \rightarrow C$. Let R_{x_2} denotes the image of f for a constant x_2 , i.e., the set $R_{x_2} = f(A, x_2) = \{y : y = f(x_1, x_2), x_1 \in A\}$. The function f is called invertible with respect to the first argument if there exists a function f_1 such that for any $x_2 \in B, y \in R_{x_2}$ $x_1 = f_1(y, x_2) \Rightarrow y = f(x_1, x_2)$. The function f_1 is called the inverse of $f : A \times B \rightarrow C$ with respect to the first argument. If for any $x_2 \in B$ the map $A \rightarrow C$ defined by $y = f(x_1, x_2)$ is surjective and injective then $y = f(x_1, x_2)$ is invertible with respect to the first variable. But even if $y = f(x_1, x_2)$ is not a bijection for a certain x_2 it can be invertible with respect to the first argument. For instance, if $y = f(x_1, x_2) = x_1^{x_2}$, $x_1 \in (0, 1]$, $x_2 \in [0, 1]$ then $x_1 = y^{1/x_2}$ for $x_2 \neq 0$, $x_1 = 1$ for $x_2 = 0$ is the inverse of f with respect to the first argument.

Maps considered here are represented by two-input Takagi-Sugeno-Kang fuzzy models with a constant consequent part of each rule. We assume that:

- the inputs are normalized such that $x_j \in [0, x_{j_{max}}]$, $j = 1, 2$,
- $m > 1$ triangular, symmetric, uniformly distributed membership functions $\mu_{j,k}(x_j)$ for which $\sum_{k=1}^m \mu_{j,k}(x_j) = 1$ are defined for each input $j = 1, 2$, as presented in Fig. 1,
- the model contains $N = m^2$ rules, the i -th rule is:
IF x_1 *is* $\mu_{1,k_1}(x_1)$ *AND* x_2 *is* $\mu_{2,k_2}(x_2)$ *THEN* $y = q_i$, $q_i = \text{const}$,
 and $Q := [q_1, \dots, q_N]$, represents the vector of consequent parameters,
- the firing strength of the i -th rule is given by the product of the input membership values: $\mu_i = \mu_{1,k_1}(x_1) \cdot \mu_{2,k_2}(x_2)$ for $1 \leq k_1, k_2 \leq m$,
- the model output is calculated as $y = \sum_{i=1}^N \mu_i q_i / \sum_{i=1}^N \mu_i$.

The schema of the considered fuzzy system is presented in Fig. 1. A fuzzy model defined in this way can be generated e.g. in Matlab using the *genfis* command with the *GridPartition* option. Only parameters q_i are tuned to match the model to the template. For example, the *lsqlin* function, which solves the least squares problem, can be used.

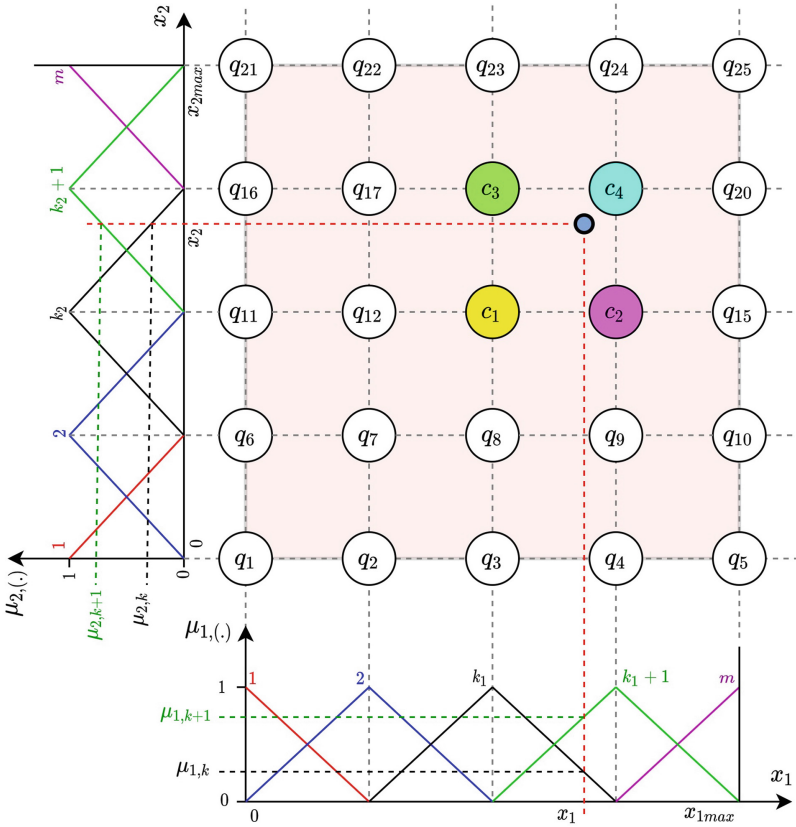


Fig. 1. Distribution of membership functions and diagram of rules

Such distribution of membership functions causes that at most two adjacent membership functions for each input are positive, and so, at most four rules can be activated (i.e. demonstrate positive firing strength $\mu_i > 0$) at the same time. We can calculate the output as a weighted sum of only four components:

$$\begin{aligned}
 y = \frac{\sum_{i=1}^N \mu_i q_i}{\sum_{i=1}^N \mu_i} &= \sum_{i=1}^N \mu_i q_i = \mu_{1,k_1}(x_1) \cdot \mu_{2,k_2}(x_2) \cdot c_1 + \mu_{1,k_1+1}(x_1) \cdot \mu_{2,k_2}(x_2) \cdot c_2 \\
 &\quad + \mu_{1,k_1}(x_1) \cdot \mu_{2,k_2+1}(x_2) \cdot c_3 + \mu_{1,k_1+1}(x_1) \cdot \mu_{2,k_2+1}(x_2) \cdot c_4,
 \end{aligned}
 \tag{1}$$

where c_i are elements of the vector Q (notice that for the presented fuzzy system $\sum_{i=1}^N \mu_i = \sum_{k_1=1}^m \sum_{k_2=1}^m \mu_{1,k_1}(x_1) \cdot \mu_{2,k_2}(x_2) = 1$). If the consequent parameters Q are organized in the vector $Q = [q_i]$ (as in Fig. 1), we have $c_1 = q_{k_1+(k_2-1)m}$, etc.

The fuzzy model defined in this way is fully determined by $N = m^2$ element vector Q and the ranges of the inputs $[0, x_{1max}]$ and $[0, x_{2max}]$. The distance between the maximum points of neighbouring (equally distributed, triangular) membership functions equals $\Delta_1 = \frac{x_{1max}}{m-1}$, and $\Delta_2 = \frac{x_{2max}}{m-1}$, for input x_1 and x_2 respectively.

Let us notice that the execution of the presented model is very fast. Knowing the input x_j , it is easy to identify two active, adjacent membership functions μ_{j,k_j} , μ_{j,k_j+1} , where

$$k_j = 1 + \text{floor}\left(\frac{x_j}{\Delta_j}\right), \quad \mu_{j,k_j} = 1 - \frac{\text{mod}(x_j - \Delta_j, \Delta_j)}{\Delta_j}, \quad \mu_{j,k_j+1} = 1 - \mu_{j,k_j}. \quad (2)$$

The selection of the appropriate rules and values $c_{(1...4)}$ is made from the set q_i based on the indices k_j determined by Eq. (2) for the inputs $j = 1, 2$. Next, four firing strengths are calculated (2 multiplications are necessary for each rule) and the final output is provided by adding 4 components. The execution time of the model is practically independent of the number of rules.

Inverted Model. The fuzzy model defined by $\{Q, \Delta_1, \Delta_2\}$, i.e. the function $(x_1, x_2) \rightarrow y = \sum_{i=1}^N \mu_i q_i$ can be inverted to determine x_1 based on the desired y for a given x_2 .

Rearranging Eqs. 1 and 2 we get

$$\begin{aligned} y &= \mu_{1,k_1} [\mu_{2,k_2} \cdot c_1 + \mu_{2,k_2+1} \cdot c_3] + \mu_{1,k_1+1} [\mu_{2,k_2} \cdot c_2 + \mu_{2,k_2+1} \cdot c_4] \\ &= \mu_{1,k_1} [\mu_{2,k_2} \cdot c_1 + \mu_{2,k_2+1} \cdot c_3] + [1 - \mu_{1,k_1}] [\mu_{2,k_2} \cdot c_2 + \mu_{2,k_2+1} \cdot c_4] \\ &= \mu_{1,k_1} [\mu_{2,k_2} \cdot c_1 - \mu_{2,k_2} \cdot c_2 + \mu_{2,k_2+1} \cdot c_3 - \mu_{2,k_2+1} \cdot c_4] + \mu_{2,k_2} \cdot c_2 + \mu_{2,k_2+1} \cdot c_4. \end{aligned} \quad (3)$$

Therefore, if k_1, k_2 are known, the value of the membership function for the first input, corresponding to given y and x_2 , is given by:

$$\mu_{1,k_1} = \frac{y - \mu_{2,k_2} \cdot c_2 - \mu_{2,k_2+1} \cdot c_4}{\mu_{2,k_2} \cdot c_1 - \mu_{2,k_2} \cdot c_2 + \mu_{2,k_2+1} \cdot c_3 - \mu_{2,k_2+1} \cdot c_4} \quad (4)$$

and so, the appropriate value of x_1 , which activates also the membership function μ_{1,k_1+1} , is:

$$x_1 = \Delta_1(k_1 - \mu_{1,k_1}). \quad (5)$$

Hence, the procedure inverting the fuzzy model (1) can be proposed:

1. Knowing y and x_2 identify k_2 and calculate $\mu_{2,k_2}(x_2)$, $\mu_{2,k_2+1}(x_2)$.

2. For $k_1 = 1, 2, \dots, m - 1$:
 identify the 4 activated rules and parameters c_1, \dots, c_4 ,
 calculate μ_{1,k_1} from equation (4),
 if $0 < \mu_{1,k_1} \leq 1$ calculate x_1 from (5), store the actual k_1, x_1 in the set Π ,
end for.
3. If the set Π of stored k_1, x_1 is empty then terminate: y is outside the range of the model for the given x_2 .
4. If the set Π contains several pairs k_1, x_1 then select the one minimizing the additional criterion (e.g. the smallest $|x_1|$, the closest to the recently observed, etc.).
5. If the set Π contains just one pair k_1, x_1 then it is the solution.

The proposed procedure provides numerical inversion of a fuzzy model - we do not derive the analytical form of the inverted model. The execution time of the proposed procedure is remarkably short and the point is that it can be easily executed on-line in a signal-processor-based control system.

The main features of the proposed fuzzy inversion are illustrated by the following examples.

Example 1. Consider the map $y = (x_2 + 2)^{(x_1+1)}$, $[x_1, x_2] \in [0, 1] \times [0, 2]$, possessing the inverse with respect to the first argument $x_1 = \log_{x_2+2}(y) - 1$. The function was represented by 32^2 uniformly distributed training triples. Several fuzzy models with different number of membership functions m were trained, and subsequently, each of them was inverted for 450 accessible pairs (y, x_2) and the calculated x_1 was compared with $\log_{x_2+2}(y) - 1$. The training error of the fuzzy model and the inversion error are presented in Table 1. The number of points is different while calculating both errors, so comparing them is not informative, but higher model accuracy provides smaller inversion errors. The ratio of the errors is more or less constant: $4.6 < RMSE/RMSE_{inv} < 8.2$.

Table 1. Root-mean-square error for the fuzzy model of $y = f(x_1, x_2)$ ($RMSE$) and for the numerical inversion ($RMSE_{inv}$), for different number of model rules $N = m^2$

m	3	5	7	10	14	21
$RMSE [y]$	0.0990	0.0248	0.0110	0.0047	0.0021	0.0005
$RMSE_{inv} [x_1]$	0.0123	0.0030	0.0013	0.0006	0.0003	0.0001

Example 2. Consider $y(x_1, x_2) = (x_1 + x_2)\sin(5x_1 + 2x_2)$, $[x_1, x_2] \in [0, 1] \times [0, 1]$ which is not a one-to-one function for some x_2 . The fuzzy model with $N = 10^2$ rules was trained using 1024 uniformly distributed training data, providing finally root-mean-square training error about 0.010. Let us denote this model output as $y_F(x_1, x_2)$.

The procedure of numerical inversion was implemented as a Simulink model. The Simulink Coder was used to generate the executable code for dSPACE1006,

and was easily executed with a sampling time $100 \mu s$, despite such a high number of rules.

The inversion was tested for the input trajectories $x_1 = 0.5(1 - \cos(2\pi t))$, $x_2 = 0.5(1 - \cos(6\pi t))$ and the corresponding $y(t)$. The time-histories are presented in Fig. 2b and as the black trajectory in Fig. 2a. For certain pairs x_2, y there exist two values x_1 generating the same y . In this case the inversion output is selected as x_1 which provides the smallest value of $J = |y(x_1, x_2) - y_F(x_1, x_2)| + \varepsilon|x_1|$. Let us denote the inverse procedure output as $x_{1F}(y, x_2)$. The experiment was repeated twice: in case A $\varepsilon > 0$ aiming for minimal $|x_{1F}|$, and in case B $\varepsilon > 0$ aiming for maximal $|x_{1F}|$. In both cases, the original input $x_1(t)$ is not reconstructed by the fuzzy inverse in the area where the mapping is not one-to-one (Fig. 2c). Despite this, the trajectory $y(t)$ is exactly reproduced by $y_F(x_{1F}(t), x_2(t))$ as it is demonstrated in Fig. 2d. The trajectory $[x_{1F}(t), x_2(t), y_F(x_{1F}(t), x_2(t))]$ moves on the surface $[x_1(t), x_2(t), y(t)]$ as it is presented in Fig. 2a.

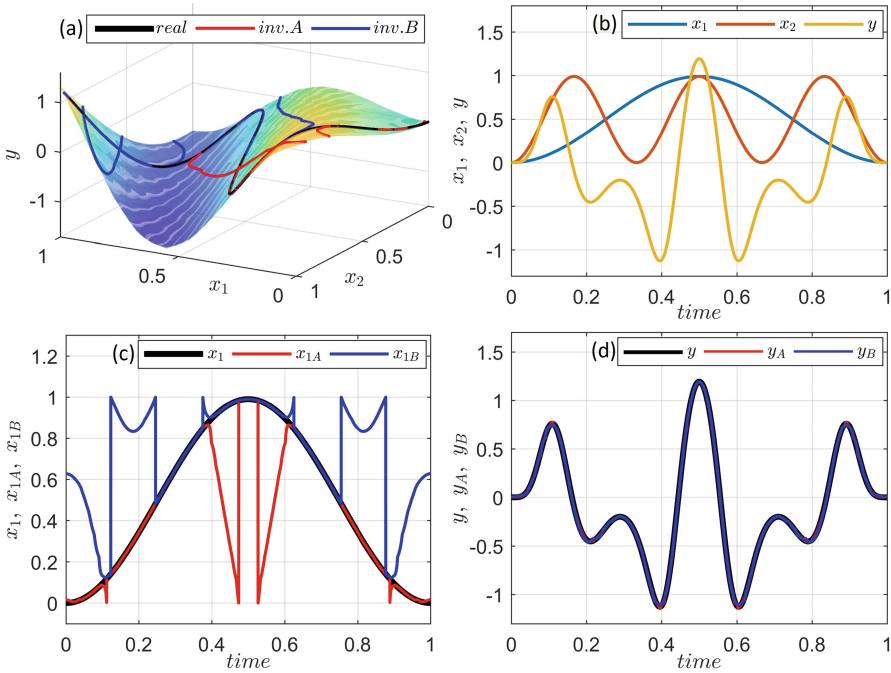


Fig. 2. The training data $y(x_1, x_2)$ and the fuzzy-modeled surface with the original and the reconstructed trajectories (a); input and output signals (b); the original and the reconstructed input $x_1(t)$ (c); the original and the reconstructed output $y(t)$

3 Linear Drive Control

Let us consider a linear drive propelled by a permanent magnet synchronous linear motor. The motion is described by differential equations:

$$\begin{aligned} \dot{x} &= v \\ M\dot{v} &= -g(x, v) + f(u, x). \end{aligned} \tag{6}$$

The state variables are position x and velocity v . M represents the mas of the drive. The propelling force f is a nonlinear function of the control u (the motor current adjusted by an inverter) and the position x . This nonlinearity is caused by differences between nonidentical permanent magnets distributed along the motor rod, eddy currents and saturation effecting the coils. Information about the shape of the surface $f = f(u, x)$ is given by the set T of 1681 uniformly distributed triples $[u, x, f]$ obtained experimentally. For the simulation we assume that

$$f(u, x) = 39 \tanh\left(\frac{1}{2}u(1+x)\right), \tag{7}$$

but the analytical form of $f = f(u, x)$ is not known.

The function $g(x, v)$ represents non-linear, speed- and position-dependent resistances to motion resulting from friction and the active external load. We assume that the mass $M = 8[kg]$ is moving in the range $-0.4 < x < 0.4[m]$ in the presence of the load

$$g(x, v) = bv + c \sin(2\pi x), \tag{8}$$

where $b = 24[Ns/m]$ and $c = 3[N]$. The nominal values of parameters M, b, c are not known, just some initial estimations can be used. The both surfaces (for nominal values of parameters) are shown in Fig. 3.

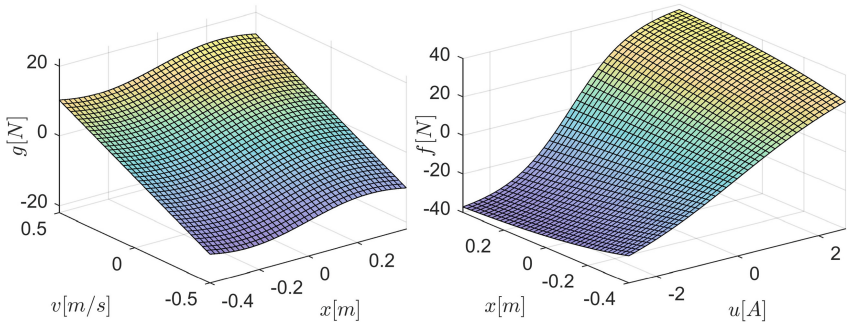


Fig. 3. Surfaces corresponding to the relationship $g(x, v)$ and $f(u, x)$

The aim of the control is to track the reference trajectory $x_d = 0.4 \sin(t)$. We start with modelling the function $f(u, x)$, on the basis of the training set T , by a

fuzzy model. To construct such a model the inputs are normalized to $[0, 0.8]$ for the shifted position and $[0, 5]$ for the shifted current. The tuning was carried out several times for a different number of rules $N = m^2$. The parameters Q of the model were selected by the least squares method based on 1681 evenly distributed training triples obtained from the training set T . Using the obtained fuzzy model, $f(u, x)$ is represented as $f(u, x) = \sum_{i=1}^N \mu_i(u, x)q_i = \mu^T(u, x)Q - \Delta(u, x)$, where $\Delta(u, x)$ denotes the modeling error, $\mu(u, x) = [\mu_1(u, x), \dots, \mu_N(u, x)]^T$ are the rules' firing strengths, and $Q = [q_1, \dots, q_N]^T$ are "the best" model parameters, ensuring the smallest possible modeling error for the given number of rules. It is assumed that the modeling error is bounded: $|\Delta(u, x)| < \delta$.

It is not necessary to know "the best" parameters exactly. Instead, the estimates $\hat{Q} = [\hat{q}_1, \dots, \hat{q}_N]$ are used in the controller. The gap between "the best" and the actual parameters is denoted by $\tilde{Q} = Q - \hat{Q}$. Finally, where $e_x = x_d - x$ is the position tracking error, $e_v = v_d - v$ is the velocity tracking error.

The backstepping method is used to design the control algorithm. Selecting the desired velocity as

$$v_d = \dot{x}_d + k_x e_x, \tag{9}$$

where $k_x > 0$ is a design parameter, provides

$$\dot{e}_x = \dot{x}_d - v = -k_1 e_x + e_v \tag{10}$$

and

$$\begin{aligned} M\dot{e}_v &= M\dot{v}_d - M\dot{v} = M\dot{v}_d + bv + c\sin(2\pi x) - \mu^T(u, x)\tilde{Q} - \mu^T(u, x)\hat{Q} + \Delta(u, x) \\ &= \phi^T\theta - \mu^T(u, x)\tilde{Q} - \mu^T(u, x)\hat{Q} + \Delta(u, x), \end{aligned} \tag{11}$$

where $\phi = [\dot{v}_d, v, \sin(2\pi x)]^T$ and $\theta = [M, b, c]^T$. The Lyapunov function

$$V = \frac{1}{2} \left[e_x^2 + M e_v^2 + \tilde{\theta}^T \Gamma_\theta^{-1} \tilde{\theta} + \tilde{Q}^T \Gamma_Q^{-1} \tilde{Q} \right] \tag{12}$$

is considered. The derivative of (12) along trajectories of (10) and (11) is

$$\dot{V} = -k_x e_x^2 + e_x e_v + e_v (\phi^T \theta - \mu^T \tilde{Q} - \mu^T \hat{Q} + \Delta) - \tilde{\theta}^T \Gamma_\theta^{-1} \dot{\tilde{\theta}} - \tilde{Q}^T \Gamma_Q^{-1} \dot{\tilde{Q}}. \tag{13}$$

Therefore, if the control u is calculated by inverting the fuzzy model $\mu^T(u, x)\hat{Q}$ so that

$$\mu^T(u, x)\hat{Q} = k_v e_v - \phi^T \hat{Q} - e_x \tag{14}$$

and the adaptive laws are

$$\dot{\tilde{\theta}} = e_v \Gamma_\theta \phi, \quad \dot{\tilde{Q}} = -e_v \Gamma_Q \mu, \tag{15}$$

then, the Lyapunov function derivative is simplified to:

$$\dot{V} = -k_x e_x^2 - (k_v - \frac{1}{2})e_v^2 - \frac{1}{2}e_v^2 + e_v \Delta. \tag{16}$$

Using the well-known inequality $e_v \Delta \leq \frac{1}{2}(e_v^2 + \Delta^2)$ and denoting $k_{min} = \min \{k_x, k_v - \frac{1}{2}\}$ we get

$$\dot{V} \leq -k_{min} \|[e_x, e_v]\|^2 + \frac{1}{2}\delta^2. \tag{17}$$

Therefore, the Lyapunov function derivative is negative outside the compact set $D = \left\{ [e_x, e_v] : \|[e_x, e_v]\|^2 \leq \frac{\delta^2}{2k_{min}} \right\}$ and this proves the stability of the system in the UUB sense [8], i.e. the trajectories $[e_x, e_v]$ ultimately approach the set D which radius can be reduced by increasing k_{min} .

As the purpose of this paper is to present the main concept of using fuzzy inverse in adaptive, nonlinear control, the simplest backstepping approach was derived. The derivation can be easily extended by introducing robust adaptive laws (with so-called σ -modification, $e - \sigma$ -modification or projection [9,10]).

The proposed algorithm was tested by numerous simulations. First, the perfect case where all parameters are known exactly was investigated. The adaptation was blocked. The modeling gap Δ is the only reason of the quasi-steady state tracking error. Table 2 presents the RMSE tuning errors of the fuzzy models and the corresponding errors of tracking the reference trajectory - the amplitude of the tracking error after reaching the quasi-steady state.

Table 2. Off-line model tuning errors and amplitude of the tracking error after reaching the quasi-steady state

m	3	5	7	10	14	21
RMSE	2.957	0.564	0.204	0.086	0.040	0.028
$max(e_x)[m]$	0.301	0.064	0.019	0.004	0.002	0.001

Next, the control system was started again to check how the adaptation of parameters $\hat{\theta}$ affects the level of tracking error resulting from the inaccuracy of the fuzzy model. The design parameters Γ_θ responsible for the speed of adaptation must be decided first, but the system is not very sensitive to this choice. The systems' performance after reaching the quasi-steady state are presented in Table 3. The adaptation resulted in a noticeable reduction in the tracking error, even though it was not directly related to the model of the generated force, nor the modeling gap.

Table 3. Amplitude of the tracking error after reaching the quasi-steady state [m]

m	3	5	7	10	14	21
without adaptation	0.301	0.064	0.019	0.006	0.002	0.001
with adaptation of M, b, c	0.046	0.018	0.007	0.004	0.001	0.0003

Finally, the effect of on-line adaptation of parameters $\hat{\theta}$ and \hat{Q} was investigated. At the start of the control algorithm, initial estimates of parameters M, b, c ranged from 50% to 150% of their true values. The off-line-tuned fuzzy model containing $N = 6^2$ rules was used to determine the control u (current) by fuzzy model inversion described in Sect. 3. The RMSE tuning error of the model was $0.32[N]$. For the first 100s of the run, only adaptation of parameters $\hat{M}, \hat{b}, \hat{c}$ has been allowed, after that time, adaptations of model parameters \hat{q}_i has started. The obtained results are shown in Figs. 4 and 5.

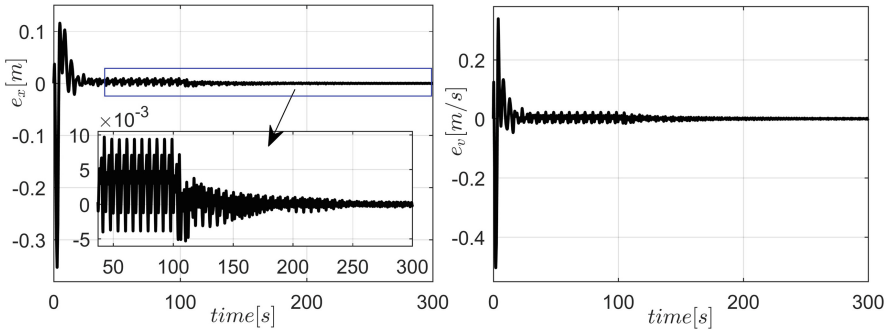


Fig. 4. Position (left) and velocity (right) trajectory tracking errors

The process of adapting the parameters $\hat{M}, \hat{b}, \hat{c}$ lasts about 30s. After this time, a quasi-steady state occurs, the amplitudes of oscillations of the errors e_x, e_v do not change and do not exceed $0.01[m]$. Small changes in the estimates (Fig. 5a) that can be observed, result from an attempt to compensate for the inaccuracy of the fuzzy model. Turning on the adaptation of \hat{q}_i at $t = 100[s]$ starts the tuning process. As a result we achieve the error of tracking the reference trajectory not bigger than $0.0003[m]$.

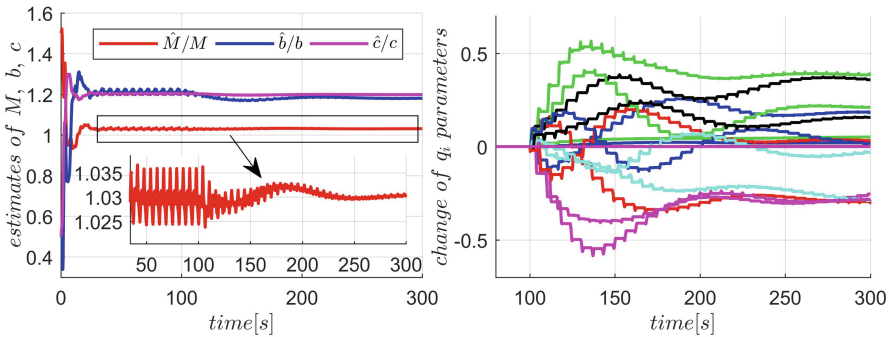


Fig. 5. Estimates of plant parameters related to their real values (left) and adaptation of fuzzy model parameters \hat{q}_i (right)

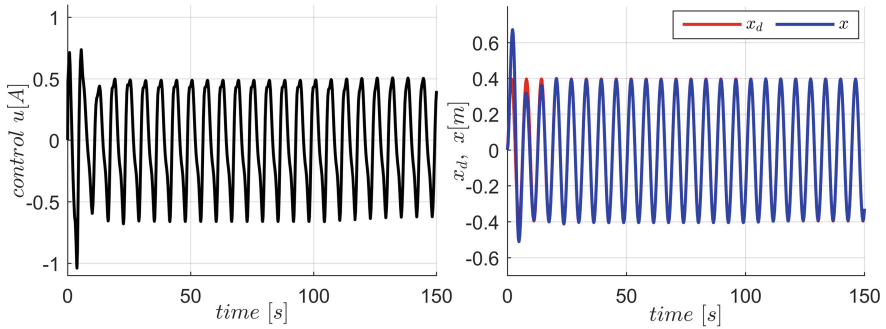


Fig. 6. The control u (left), the desired trajectory x_d and the actual position x

The control (the motor current) is presented in Fig. 6. It is smooth, bounded and oscillates according to the desired trajectory.

Simulations were repeated with the fuzzy model using $N = 21^2$ rules. The tracking accuracy was improved - tracking errors without the on-line adaptation of \hat{q}_i were similar as in the case of $N = 6^2$ rules with the on-line adaptation.

4 Conclusions

The presented concept of non-linear adaptive control can be used for any system with complex, nonlinear dependencies of control and state variables. It can be applied even when the plant model is partially represented by digital data only, or is too complicated to use an analytical model. The proposed approach assumes modeling of such a nonlinear map $f(u, x)$ of control and state variables by a fuzzy TSK system and the special numerical inversion of the model, performed on-line to find u assuring the desired trajectory of f . Although the presented derivation was limited to two-input systems, it can be easily generalized to multi-input maps that are inverted with respect to any single argument. The proposed fuzzy model together with the derived procedure of numerical inversion offers two important advantages: first, the execution time of all necessary arithmetical operations is very short, so the presented approach can be implemented in numerous DSP-based control system, even when the sampling is very fast; second, the model parameters can be improved on-line by adaptation, hence the perfect off-line training is not necessary.

The advantages of the proposed technique are illustrated by a simple example of adaptive backstepping control of a linear drive. It was confirmed that better tracking accuracy can be achieved by adaptation, even for less accurate and simpler models resulting from off-line training.

The same approach can be incorporated into many other nonlinear control techniques, for instance, adaptive backstepping with tuning functions, adaptive nonlinear model-following control, passivity-based control.

Due to the space limit, only one numerical example is presented here. The authors can report numerous implementation experiments concerning real-life systems, mainly regarding motion control, electric drives or pneumatic servo drives, where the fuzzy inversion technique effectively copes with complex phenomena affecting the air flow through the valve. In each case, the necessary calculations were easily made within the short sampling period required for proper system operation.

The signaled generalizations of the proposed approach and subsequent applications will be reported in our future works.

References

1. Krstic, M., Kanellakopoulos, I., Kokotovic, P.V.: *Nonlinear and Adaptive Control Design*. Wiley-Interscience June 1995. ISBN: 978-0-471-12732-1
2. Liu, Z., Dong, X., Xue, J., Li, H., Chen, Y.: Adaptive neural control for a class of pure-feedback nonlinear systems via dynamic surface technique. *IEEE Trans. Neural Networks Learn. Syst.* **27**(9) (2016)
3. Triska, L., Portella, J., Reger, J.: Dynamic extension for adaptive backstepping control of uncertain pure-feedback systems, *IFAC-PapersOnLine* **54**(14), 307–312 (2021). ISSN 2405–8963, <https://doi.org/10.1016/j.ifacol.2021.10.371>
4. Peng, J., Dubay, R.: Adaptive fuzzy backstepping control for a class of uncertain nonlinear strict-feedback systems based on dynamic surface control approach. *Expert Systems with Applications* **120**, 239–252 (2019). ISSN 0957–4174, <https://doi.org/10.1016/j.eswa.2018.11.040>
5. Kabziński, J.: Adaptive, compensating control of wheel slip in railway vehicles. *Bullet. Pol. Acad. Sci. Tech. Sci.* **63**(4) (2015). <https://doi.org/10.1515/bpasts-2015-0108>
6. Ulu, C., Güzelkaya, M., Eksin, I.: Exact analytical inversion of TSK fuzzy systems with singleton and linear consequents. *Int. J. Approx. Reason.* **55**(6) (2014). <https://doi.org/10.1016/j.ijar.2014.01.007>
7. K. Siminski, Ridders algorithm in approximate inversion of fuzzy model with parametrized consequences, *Expert Syst. Appl.* **51**, 276–285 (2016). ISSN 0957–4174, <https://doi.org/10.1016/j.eswa.2015.12.042>
8. Kabziński, J., Mosiolek, P.: *Nonlinear Control Design*, 1st edn. PWN SA: Warszawa (2018). ISBN 978-83-01-19697-4
9. Ioannou, P., Sun, J.: *Robust Adaptive Control*. Dover Publicatos, Inc. Mineola, New York (2012), ISBN 978-04-86-49817-1
10. Sastry, S., Bodson, M.: *Adaptive Control: Stability, Convergence and Robustness*. Dover Publicatos, Inc. Mineola, New York (2011). ISBN 978-04-86-48202-6



On the Choice of the Cost Function for Nonlinear Model Predictive Control: A Multi-criteria Evaluation

Robert Nebeluk^(✉) and Maciej Ławryńczuk

Institute of Control and Computation Engineering, Warsaw University of Technology,
ul. Nowowiejska 15/19, 00-665 Warsaw, Poland
{Robert.Nebeluk,Maciej.Lawrynczuk}@pw.edu.pl

Abstract. Typically, Model Predictive Control (MPC) algorithms minimise the sum of squared predicted control errors, i.e., the L_2 norm. This work reviews the possible cost functions that may be used in MPC. Eight different cost functions are considered and their usefulness is investigated. For a neutralisation reactor benchmark, the influence of the cost function on the resulting control quality is compared in terms of as many as eight indices. All studied approaches are compared in two scenarios: with no process disturbances and when measurement noise and a disturbance affect the process.

Keywords: Model Predictive Control · Control quality assessment · Cost function

1 Introduction

Model Predictive Control (MPC) algorithms [12, 18] optimise on-line the control policy using a model of the process. The model is used for prediction of future process behaviour. Such a formulation gives very good control quality, in particular for multivariable and nonlinear processes. Additionally, MPC algorithms have the unique possibility of considering constraints imposed on process variables. As a result, MPC algorithms find numerous applications in different fields; a few examples can be named: servomotors [8], autonomous vehicles [4], quadrotors [5], electromagnetic mills [15], vehicle suspension systems [16] and stochastic systems [1]. It is observed that MPC algorithms that use different neural network structures are increasingly popular [10, 11, 14, 16, 17].

MPC algorithms calculate the future control policy as a result of minimisation of the predicted control errors determined on-line from the process model. The predicted control errors are typically squared, corresponding to the classical L_2 norm. It is because such a choice has good numerical properties and computationally uncomplicated quadratic optimisation tasks are solved provided that linear models are used for prediction [12, 18]. Minimisation of the sum of absolute values of the predicted control errors, i.e., the L_1 norm, is significantly less frequent.

Although such an approach is reported to give better control quality than the classical L_2 norm, e.g., [3, 6, 9, 11, 14], in the case of a nonlinear model, the L_1 norm requires nonlinear optimisation [6] or advanced on-line linearisation which leads to quadratic optimisation [11, 14]. This work reviews the possible cost functions that may be used in MPC; as many as eight options are discussed. Moreover, the usefulness of all considered cost functions is investigated in a nonlinear MPC algorithm of a neutralisation reactor benchmark. A multi-criteria comparison is performed in which the control quality is assessed in terms of as many as eight indices. All approaches are compared in two scenarios: with no process disturbances and when measurement noise and a disturbance affect the process.

2 MPC Problem Formulation

The following MPC optimisation task [18] is considered

$$\begin{aligned} & \min_{\Delta \mathbf{u}(k)} \{J(k)\} \\ & \text{subject to} \\ & u^{\min} \leq u(k+p|k) \leq u^{\max}, \quad p = 0, \dots, N_u - 1 \\ & \Delta u^{\min} \leq \Delta u(k+p|k) \leq \Delta u^{\max}, \quad p = 0, \dots, N_u - 1 \\ & y^{\min} \leq \Delta \hat{y}(k+p|k) \leq y^{\max}, \quad p = 1, \dots, N \end{aligned} \quad (1)$$

where, N_u and N stand for the control and prediction horizons, respectively. This work considers the Single Input Single Output (SISO) case for a short presentation. Hence, the constraints are defined by scalars u^{\min} , u^{\max} , Δu^{\min} , Δu^{\max} , y^{\min} , y^{\max} and the vector of decision variables contains increments of the future manipulated variable

$$\Delta \mathbf{u}(k) = [\Delta u(k|k) \dots \Delta u(k+N_u-1|k)]^T \quad (2)$$

At each discrete sampling instant, the optimisation task (1) is solved and the first element of the solution vector (2) is applied to the process.

In general, the minimised cost function can be expressed as the sum of two parts

$$J(k) = J_y(k) + J_u(k) \quad (3)$$

Typically, the first part of the cost function, $J_y(k)$, measures the future control errors on the prediction horizon. In contrast, the second one, $J_u(k)$, may weigh the values or (and) the increments of the future control policy. A very frequent choice is to consider the sum of squared increments of the calculated decision variables over the control horizon to penalise excessive increments of the manipulated variable

$$J_u(k) = \lambda \sum_{p=0}^{N_u-1} (\Delta u(k+p|k))^2 \quad (4)$$

This approach is used in this study, regardless of the choice of the first term of the MPC cost function.

3 Possible MPC Cost Functions

Let us review possible choices for the first part of the MPC cost function. As it is shown in Sect. 4, the choice of the function $J_y(k)$ has an important impact on the obtained control quality.

The classical L_2 norm measures the sum of squared control errors on the prediction horizon [12]

$$J_y(k) = \sum_{p=1}^N (e(k+p|k))^2 \tag{5}$$

where

$$e(k+p|k) = y^{sp}(k+p|k) - \hat{y}(k+p|k) \tag{6}$$

is the predicted control error for the future sampling instant $k+p$ computed at the current time step k ; $y^{sp}(k+p|k)$ and $\hat{y}(k+p|k)$ stand for the set-point and predicted values of the controlled variable, respectively.

The L_1 norm measures absolute values of the predicted control errors [12]

$$J_y(k) = \sum_{p=1}^N |e(k+p|k)| \tag{7}$$

The L_∞ norm takes into account the maximal absolute value of the predicted control error over the prediction horizon [12]

$$J_y(k) = \max(|e(k+1|k)|, \dots, |e(k+N|k)|) \tag{8}$$

The L_{Huber} norm utilises the Huber function [13]

$$J_y(k) = \sum_{p=1}^N z(k+p|k) \tag{9}$$

where

$$z(k+p|k) = \begin{cases} \frac{(e(k+p|k))^2}{2} & \text{if } |e(k+p|k)| \leq h \\ h(|e(k+p|k)| - \frac{h}{2}) & \text{if } |e(k+p|k)| > h \end{cases} \tag{10}$$

and h is the shape parameter.

The L_{Cauchy} norm relies on the Cauchy function [13]

$$J_y(k) = \sum_{p=1}^N \frac{(e(k+p|k))^2}{2} \log \left(1 + \frac{(e(k+p|k))^2}{c} \right) \tag{11}$$

where c is the Cauchy function parameter.

The $L_{\text{SC-DCS}}$ (Dynamic Covariance Scaling) norm uses Eq. (9), but now [13]

$$z(k+p|k) = \begin{cases} \frac{(e(k+p|k))^2}{2} & \text{if } (e(k+p|k))^2 \leq \phi \\ \frac{2\phi(e(k+p|k))^2}{\phi+(e(k+p|k))^2} - \frac{\phi}{2} & \text{if } (e(k+p|k))^2 > \phi \end{cases} \tag{12}$$

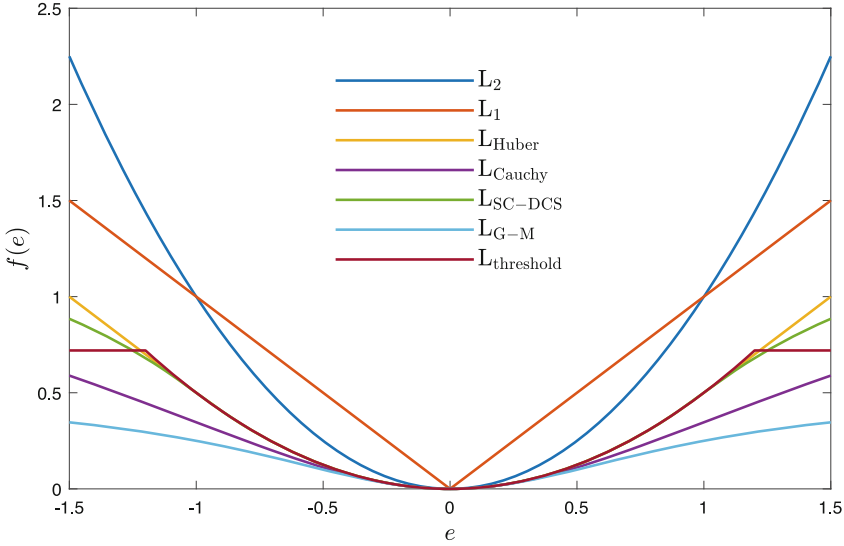


Fig. 1. Plots of the considered cost functions

where ϕ is the SC-DCS function shape parameter.

The $L_{\text{threshold}}$ norm also uses Eq. (9), but now [13]

$$z(k + p|k) = \begin{cases} \frac{(e(k+p|k))^2}{2} & \text{if } |e(k + p|k)| \leq s \\ \frac{s^2}{2} & \text{if } |e(k + p|k)| > s \end{cases} \quad (13)$$

where s is the shape parameter of the function.

The L_{G-M} (Geman-McClure) norm has the following form [13]

$$J_y(k) = \sum_{p=1}^N \frac{0.5 (e(k + p|k))^2}{\sigma + (e(k + p|k))^2} \quad (14)$$

where σ is a parameter of the G-M function. All mentioned functions, except the L_∞ norm, are shown in Fig. 1. The control error range has been selected adequately to the process described in Sect. 4.

Taking into account the cost functions defined above, the following MPC algorithms are considered: MPC- L_2 , MPC- L_1 , MPC- L_∞ , MPC- L_{Huber} , MPC- L_{Cauchy} , MPC- $L_{\text{SC-DCS}}$, MPC- L_{G-M} and MPC- $L_{\text{threshold}}$.

4 Simulation Results

In this Section, simulation results of MPC algorithms are presented and studied in which all eight cost functions reviewed in Sect. 3 are used. Precisely, the alternative cost functions defined by Eqs. (5) and (7)–(14) are used only as

the first part of the MPC cost function in Eq. (3), i.e., $J_y(k)$. The second part of the MPC cost function, i.e., $J_u(k)$, is always a penalty term which measures squared increments of future increments of the manipulated variables, as defined by Eq. (4).

All MPC algorithms with different cost functions use a neutralisation reactor benchmark process. The process has one manipulated variable (base NaOH stream q_1 (ml/s)), one disturbance variable (buffer NaHCO_3 stream q_2 (ml/s)) and one controlled variable (pH of the product). The detailed fundamental model of the process is given in [7]. The prediction for all MPC algorithms is calculated using a Wiener process model consisting of a linear dynamic block and a nonlinear static block [10]. A sigmoid-like neural network is used in the nonlinear static block. A thorough discussion of model training, validation and selection is given in [10]. Nonlinear optimisation is used to solve the MPC optimisation task (1).

Two cases are considered: the undisturbed simulation scenario and the disturbed simulation scenario. In the first case, five changes of the set-point value are considered and no disturbance influences the controlled process, i.e., the disturbance variable (buffer) is constant. The simulation horizon is rather short, as 120 sampling steps are only considered. In the second scenario, as many as 50000 sampling steps are considered and the process is affected by two disturbances. Firstly, the disturbance variable, i.e., q_2 , changes; a real industrial disturbance is applied during the simulation. Secondly, small measurement noise of the controlled variable of the process is considered. In all simulations, parameters of all the MPC algorithms are set to $N = 10$, $N_u = 3$ and $\lambda = 0.1$ [10] and remain unchanged for all considered experiments. The sampling period is 10 s. Specific parameters of the cost functions are: $h = 1$ (for the algorithm with the L_{Huber} norm (Eq. (10)), $c = 1$ (for the algorithm with the L_{Cauchy} norm (Eq. (11)), $\phi = 1$ (for the algorithm with the term (12), $\sigma = 1$ (for the algorithm with the $L_{\text{SC-DCS}}$ norm (Eq. (14)) and $s = 1.2$ (for the algorithm with the $L_{\text{threshold}}$ norm (Eq. (13)). All mentioned parameters are tuned so that the algorithms work within the control error range of the considered process.

A multi-criteria approach is used to evaluate the control performance. The following control quality indicators are calculated after simulations [2]:

- the Mean Squared Error (MSE) of the control error e ,
- the Mean Absolute Error (MAE) of the control error e ,
- the Gauss standard deviation (σ_G) of the control error e ,
- the Huber standard deviation (σ_H) of the control error e ,
- the scale factor of the alpha-stable distribution (γ_α) of the control error e ,
- the differential entropy (H_D) of the control error e ,
- the rational entropy (H_R) of the control error e ,
- the median (LMS) of the control error e .

4.1 Simulation Results: The Undisturbed Process Scenario

For the first simulation scenario, the values of eight considered indices are given in Table 1. One can observe the following:

Table 1. Comparison of control quality indices when different MPC cost functions are used: the undisturbed simulation scenario

MPC cost function	MSE	MAE	σ_G	σ_H
L_2	1.48×10^0	4.87×10^{-1}	1.22×10^0	2.38×10^{-2}
L_1	1.47×10^0	4.71×10^{-1}	1.21×10^0	1.27×10^{-2}
L_∞	1.64×10^0	5.39×10^{-1}	1.28×10^0	1.58×10^{-2}
L_{Huber}	1.53×10^0	5.17×10^{-1}	1.24×10^0	3.41×10^{-2}
L_{Cauchy}	1.54×10^0	5.22×10^{-1}	1.24×10^0	3.42×10^{-2}
$L_{\text{SC/DCS}}$	1.53×10^0	5.17×10^{-1}	1.24×10^0	3.41×10^{-2}
$L_{\text{G-M}}$	1.58×10^0	5.31×10^{-1}	1.26×10^0	3.47×10^{-2}
$L_{\text{threshold}}$	1.53×10^0	5.17×10^{-1}	1.24×10^0	3.41×10^{-2}

MPC cost function	γ_α	H_D	H_R	LMS
L_2	8.2×10^{-3}	7.89×10^0	3.26×10^{-1}	1.71×10^{-4}
L_1	5.41×10^{-3}	8.74×10^0	3.11×10^{-1}	5.46×10^{-5}
L_∞	6.58×10^{-3}	8.06×10^0	3.51×10^{-1}	9.38×10^{-5}
L_{Huber}	1.24×10^{-2}	7.13×10^0	3.61×10^{-1}	2.75×10^{-4}
L_{Cauchy}	1.06×10^{-2}	7.15×10^0	3.67×10^{-1}	2.78×10^{-4}
$L_{\text{SC/DCS}}$	1.24×10^{-2}	7.13×10^0	3.61×10^{-1}	2.75×10^{-4}
$L_{\text{G-M}}$	1.08×10^{-2}	7.05×10^0	3.82×10^{-1}	2.76×10^{-4}
$L_{\text{threshold}}$	1.24×10^{-2}	7.13×10^0	3.61×10^{-1}	2.75×10^{-4}

1. The lowest values of MSE, MAE, σ_G , σ_H , γ_α , H_R and LMS indicators are obtained for the L_1 norm.
2. The lowest value of the H_D indicator is possible for the $L_{\text{G-M}}$ norm.
3. The second lowest values of MSE, MAE, σ_G and H_R indicators are obtained for the L_2 norm.
4. The second lowest values of σ_H , γ_α and LMS indicators are found for the L_∞ norm.
5. The second lowest values of the H_D indicator is acquired for L_{Huber} , $L_{\text{SC-DCS}}$ and $L_{\text{threshold}}$ norms.
6. The largest values of MSE, MAE and σ_G indicators are obtained for the L_∞ norm.
7. The largest value of γ_α indicator is found for L_{Huber} , $L_{\text{SC-DCS}}$ and $L_{\text{threshold}}$ norms.
8. The largest value of the H_D indicator is acquired for the L_1 norm.
9. The largest value of the H_R indicator is obtained for the $L_{\text{G-M}}$ norm.
10. The largest value of the LMS indicator is found for the L_{Cauchy} norm.

Process trajectories are shown in Fig. 2. The following can be noted:

1. The L_1 norm results in fast control and minimal or no overshoot, especially for negative control errors.
2. The L_∞ norm can give better control for negative control errors, but, unfortunately, it gives large overshoot for positive control errors.
3. The L_2 norm gives the second best control quality in terms of the settling time and the overshoot.

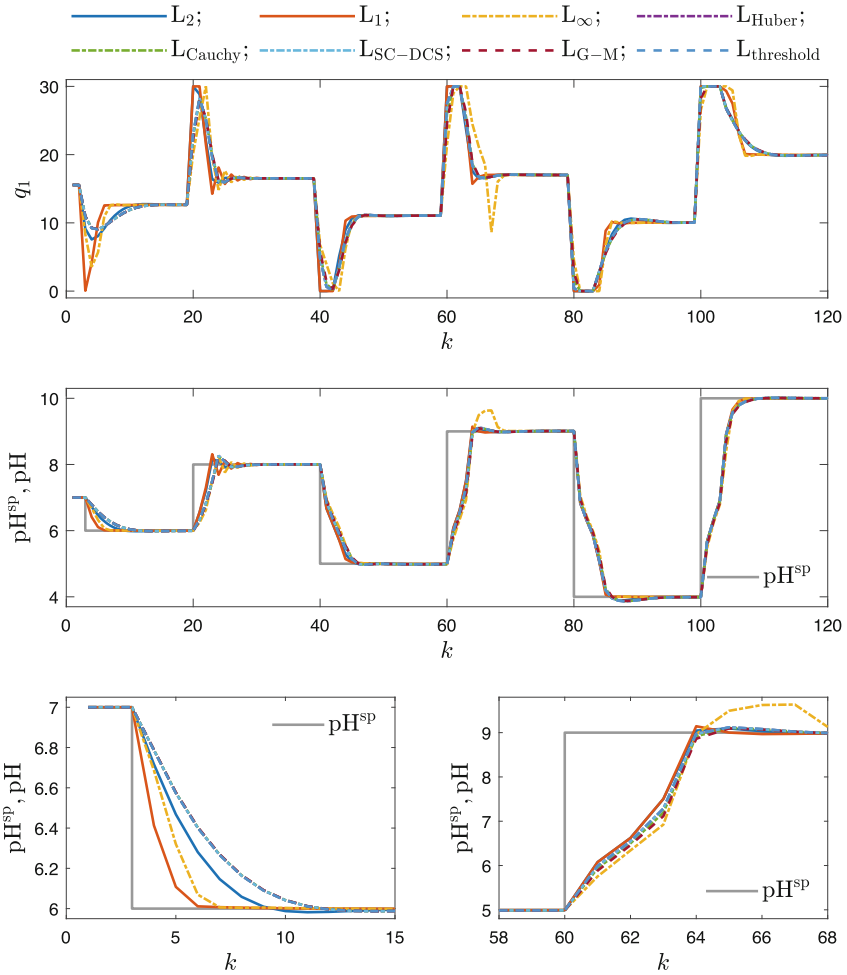


Fig. 2. Comparison of process trajectories when different MPC cost functions are used: the undisturbed simulation scenario. Two top panels show the results for the whole simulation horizon; the bottom panels show two example enlarged fragments.

4.2 Simulation Results: The Disturbed Process Scenario

For the second simulation scenario, the obtained values of eight considered indices are given in Table 2. Chosen example fragments of process trajectories are shown in Fig. 3. One can observe the following:

1. Similarly to the first simulation scenario, the lowest values of indicators MSE, MAE, σ_G and H_R are found for the L_1 norm and the lowest value of the H_D indicator is obtained for the L_{G-M} norm.
2. Unlike the previous simulations, the lowest values of σ_H , γ_α and LMS indicators are acquired for the L_2 norm.

Table 2. Comparison of control quality indices when different MPC cost functions are used: the disturbed simulation scenario

MPC cost function	MSE	MAE	σ_G	σ_H
L_2	1.25×10^{-1}	1.13×10^{-1}	3.54×10^{-1}	4.52×10^{-2}
L_1	9.60×10^{-2}	9.28×10^{-2}	3.10×10^{-1}	5.08×10^{-2}
L_∞	1.79×10^{-1}	1.41×10^{-1}	4.23×10^{-1}	5.42×10^{-2}
L_{Huber}	1.50×10^{-1}	1.31×10^{-1}	3.87×10^{-1}	4.70×10^{-2}
L_{Cauchy}	1.52×10^{-1}	1.32×10^{-1}	3.90×10^{-1}	4.71×10^{-2}
$L_{\text{SC/DCS}}$	1.51×10^{-1}	1.31×10^{-1}	3.87×10^{-1}	4.70×10^{-2}
$L_{\text{G-M}}$	1.55×10^{-1}	1.33×10^{-1}	3.93×10^{-1}	4.72×10^{-2}
$L_{\text{threshold}}$	1.50×10^{-1}	1.31×10^{-1}	3.87×10^{-1}	4.70×10^{-2}

MPC cost function	γ_α	H_D	H_R	LMS
L_2	2.76×10^{-2}	7.61×10^3	2.63×10^0	9.10×10^{-4}
L_1	3.47×10^{-2}	7.71×10^3	2.56×10^0	1.16×10^{-3}
L_∞	3.51×10^{-2}	7.49×10^3	2.75×10^0	1.27×10^{-3}
L_{Huber}	2.83×10^{-2}	7.47×10^3	2.67×10^0	9.68×10^{-4}
L_{Cauchy}	2.83×10^{-2}	7.47×10^3	2.67×10^0	9.68×10^{-4}
$L_{\text{SC/DCS}}$	2.83×10^{-2}	7.47×10^3	2.67×10^0	9.68×10^{-4}
$L_{\text{G-M}}$	2.83×10^{-2}	7.46×10^3	2.67×10^0	9.71×10^{-4}
$L_{\text{threshold}}$	2.83×10^{-2}	7.47×10^3	2.67×10^0	9.68×10^{-4}

3. Similarly to the first simulation scenario, the largest values of:
 - (a) MSE, MAE and σ_G indicators are obtained for the L_∞ norm,
 - (b) the σ_H indicator is found for the $L_{\text{G-M}}$ norm,
 - (c) the γ_α indicator is acquired for L_{Huber} , $L_{\text{SC-DCS}}$ and $L_{\text{threshold}}$ norms,
 - (d) the H_D indicator is obtained for the L_1 norm,
 - (e) the H_R indicator is found for the $L_{\text{G-M}}$ norm,
 - (f) the LMS indicator is acquired for the L_{Cauchy} norm.
4. Trajectories presented in Fig. 3 give the same conclusions as those for the first simulation scenario, i.e., the L_1 norm performs best.

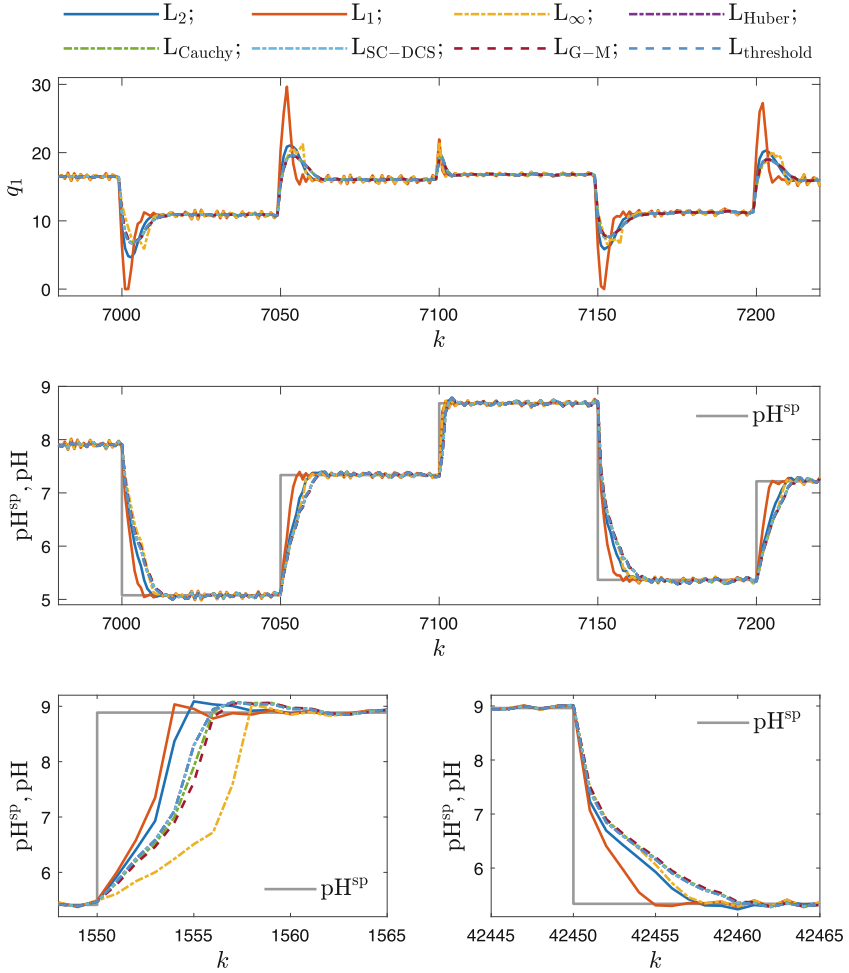


Fig. 3. Comparison of process trajectories when different MPC cost functions are used: the disturbed simulation scenario. Two top panels show the results for the whole simulation horizon; the bottom panels show two example enlarged fragments.

5 Conclusions

This paper presents a detailed investigation of the impact of different MPC cost functions on the resulting control performance in the control system of a simulated nonlinear neutralisation reactor. Typically used MPC cost functions are discussed, i.e., L_1 , L_2 and L_∞ norms, and a few alternatives such as L_{Huber} , L_{Cauchy} , $L_{\text{SC-DCS}}$, $L_{\text{G-M}}$ and $L_{\text{threshold}}$ norms. Two simulation scenarios are considered: undisturbed and disturbed process cases; real industrial disturbances are utilised in the latter case. Overall, in both cases, the MPC algorithm with the L_1 norm gives the best results, while the typically used L_2 norm leads to

the second best results. All other norms give worse control quality, regardless of which control quality indices are used. It is because the L_1 norm considers the predicted control errors linearly, while other norms are sensitive to large errors but tend to overlook smaller ones. In future works, the authors consider developing efficient MPC algorithms for some of the discussed norms.

Acknowledgement. This research was financed by Warsaw University of Technology in the framework of the project for the scientific discipline automatic control, electronics and electrical engineering.

References

1. Bania, P.: An information based approach to stochastic control problems. *Int. J. Appl. Math. Comput. Sci.* **30**, 23–34 (2020)
2. Domański, P.: Control Performance Assessment: Theoretical Analyses and Industrial Practice, *Studies in Systems, Decision and Control*, vol. 245. Springer, Cham (2020). 0.1007/978-3-030-23593-2
3. Domański, P., Ławryńczuk, M.: Impact of MPC embedded performance index on control quality. *IEEE Access* **9**, 24787–24795 (2021)
4. Ducajů, S., Salt Llobregat, J.J., Cuenca, Á., Tomizuka, M.: Autonomous ground vehicle lane-keeping LPV model-based control: dual-rate state estimation and comparison of different real-time control strategies. *Sensors* **21**, 1531 (2021)
5. Eskandarpour, A., Sharf, I.: A constrained error-based MPC for path following of quadrotor with stability analysis. *Nonlinear Dyn.* **98**, 899–918 (2020)
6. Fehér, M., Straka, O., Šmídl, V.: Model predictive control of electric drive system with L_1 -norm. *Eur. J. Control* **56**, 242–253 (2020)
7. Gómez, J.C., Jutan, A., Baeyens, E.: Wiener model identification and predictive control of a pH neutralisation process. *Proc. IEE, Part D, Control Theory Appl.* **151**, 329–338 (2004)
8. Horla, D.: Experimental results on actuator/sensor failures in adaptive GPC position control. *Actuators* **10**(3), 1–18 (2021)
9. Karamanakos, P., Geyer, T., Kennel, R.: On the choice of norm in finite control set model predictive control. *IEEE Trans. Power Electron.* **33**(8), 7105–7117 (2018)
10. Ławryńczuk, M.: Nonlinear Predictive Control Using Wiener Models: Computationally Efficient Approaches for Polynomial and Neural Structures, *Studies in Systems, Decision and Control*, vol. 389. Springer, Cham (2022). <https://doi.org/10.1007/978-3-030-83815-7>
11. Ławryńczuk, M., Nebeluk, R.: Computationally efficient nonlinear model predictive control using the L_1 cost-function. *Sensors* **21**(17) (2021)
12. Maciejowski, J.: *Predictive Control with Constraints*. Prentice Hall, Harlow (2002)
13. MacTavish, K., Barfoot, T.D.: At all costs: a comparison of robust cost functions for camera correspondence outliers. In: 2015 12th Conference on Computer and Robot Vision. pp. 62–69. Halifax, NS, Canada (2015)
14. Nebeluk, R., Ławryńczuk, M.: Fast model predictive control of PEM fuel cell system using the L_1 norm. *Energies* **15**(14) (2022)
15. Ogonowski, S., Bismor, D., Ogonowski, Z.: Control of complex dynamic nonlinear loading process for electromagnetic mill. *Arch. Control Sci.* **30**, 471–500 (2020)
16. Papadimitrakis, M., Alexandridis, A.: Active vehicle suspension control using road preview model predictive control and radial basis function networks. *Appl. Soft Comput.* **120**, 108646 (2022)

17. Schwedersky, B.B., Flesch, R.C.C.: Nonlinear model predictive control algorithm with iterative nonlinear prediction and linearization for long short-term memory network models. *Eng. Appl. Artif. Intell.* **115**, 105247 (2022)
18. Tatjewski, P.: *Advanced Control of Industrial Processes, Structures and Algorithms*. Springer, London (2007). <https://doi.org/10.1007/978-1-84628-635-3>



Adaptive Sliding Mode control of Traffic Flow in Uncertain Urban Networks

Ali Soltani Sharif Abadi¹(✉), Pooyan Alinaghi Hosseinabadi², and Andrew Ordys¹

¹ Institute of Automatic Control and Robotics, Faculty of Mechatronics,
Warsaw University of Technology, Warsaw, Poland
ali.soltani_sharif_abadi.dokt@pw.edu.pl

² School of Engineering and Information Technology, The University of New South Wales,
Canberra, ACT, Australia

Abstract. One of the most significant issues in the cities is the heavy traffic that everyday citizens are involved with. In recent decades, various sciences, such as computing science and control and traffic science have been incorporated to address this problem. In this paper, theoretically, the Terminal Sliding Mode Control (TSMC) method is proposed to control the traffic flow in urban networks. Firstly, a two regions model for urban traffic with various uncertainties and external disturbances is considered; subsequently, the Lyapunov stability method is employed to control the urban traffic flow in this model by using designed control inputs and adaptive laws. Finally, a numerical simulation for a case study is performed in Simulink/MATLAB to reveal the effectiveness of the proposed designs in this paper.

Keywords: Traffic Control · Adaptive · Terminal Sliding Mode Control · Finite-time · Urban Networks

1 Introduction

The urban road networks are heavily congested in many countries, which results in some unwanted issues such as increased travel costs, increased travel times, increased number of stops, unavoidable delays, increased air pollution and noise pollution, increased distress for drivers and passengers, and increased the number of road accidents [1, 2]. Increasing the capacity of traffic networks by constructing more roads is very costly and can damage the environment. Traffic engineers are constantly working to make a well-organized transportation system that aims to travel demand management by using transport infrastructure and traffic control strategies [3]. Designers, engineers, and economists have improved and implemented various travel demand management methodologies to decrease the fast-increasing traffic congestion. The main methods for travel demand management are listed as follows: utilizing park and ride, a particular lane for passing heavy vehicles, using a sharing vehicle, paying tolls at the city's central area entrance, parking charges for a vehicle and so on. These strategies might reduce the demand for the existing transportation system [4–6].

Traffic engineers also continuously make an effort to minimize the total time delay, the number of stopped vehicles, and the release of harmful emissions to the environment by using different traffic strategies and traffic management methodologies. Traffic signals facilitate pedestrians to cross the road, help the stopped vehicles at the side of the street go back to the direct traffic, and make it easier for progressive traffic flow of the main streets. The engineering practice demonstrates that many traffic signals can be enhanced by updating scheduling programs [7, 8].

Urban traffic control systems consist of three developed generations. Based on past data considered the first generation, the systems have been worked on. The second generation has surpassed the first generation because of detectors, which can collect a set of traffic data in real-time to rearrange and select traffic signal programs. The third generation provides the ability to predict traffic conditions. It provides plans for traffic signals programs and approaches pre-computed and applied at the proper time for optimal control of current traffic conditions. The fourth generation of traffic control systems is currently under upgrading, which has been developed based on the principles of artificial intelligence and the ability to provide information at any time by traffic forecasting and accident identification based on engineering principles for integrated systems on a large scale. Although these systems mainly utilize advances in various information technology, it is clear that their performance is always related to optimization methods and essential control approaches [9, 10].

It has been stated in [11] that an appropriate urban traffic network model is necessary to control the urban traffic network by optimization methods. This model should be precise in describing network traffic as well as simple to compute. Firstly, a link model has been provided to describe dynamic traffic behaviors, improving accuracy. Subsequently, a general urban traffic network has been proposed. Accordingly, a macroscopic model has been created for urban traffic networks by modeling the network elements. Also, a comparison has been made between the proposed model in [11] and the microscopic traffic model (CORSIM), which has demonstrated that the proposed model has a good balance in terms of accuracy and simpleness. Hence it is suitable to control in real time. The uncertainty of the modeling and constrained robust urban traffic control has been investigated [12]. The linear polytopic method has been used to describe the uncertain network system. Also, the robust and infinite horizon model predictive control (MPC) methodology has been proposed to cope with model mismatches. In [13], the efficiency of the signal control strategy has been investigated. Indeed, the problem of critical congested links has been addressed, which is implementable in a control structure in large-scale, heterogeneous urban traffic networks.

The model predictive control (MPC) methodology with possible constraints has been introduced in [14] to control the traffic signal in urban traffic networks. In this research, the uncertainty of the inflow to the traffic network has been mainly addressed based on the MPC strategy. A dynamic routing methodology has been proposed in [15] by considering adaptive signal control for efficient routing in real-time traffic networks. Indeed, from a practical point of view, choosing the route of travelers can be affected by traffic control methodology in the transportation networks. The interaction between choosing a route by travelers and traffic signal control has been considered in a framework.

The Huangshan Road in Hefei, China, has been considered in [16] to evaluate developing an adaptive signal control system utilizing the VS-PLUS strategy for optimal network traffic control. This research included improving the trip matrix utilizing an enhanced Furness methodology and field study about traffic performance. Perimeter traffic flow control has been improved in [17] based on Macroscopic Fundamental Diagram (MFD) to control traffic congestion in heterogeneously congested cities. Robust control for the two-region MFD system has been proposed to deal with various uncertainties and external disturbances. Indeed, the Sliding Mode Control (SMC) methodology has been employed due to its main feature: robustness against various uncertainties. The Linear Quadratic Regulator (LQR) control strategy has also been designed to evaluate the SMC scheme further.

Nowadays, pre-described time control methods are developing. The concepts of these methods are presented in [18, 19]. The Finite-time control method and stability is one the most popular control pre-described time methods [20]. The Fixed-time [21–24] and predefined-time [25, 26] are the newest control methods in this field.

This paper proposes the Terminal Sliding Mode Control (TSMC) method to design control inputs for a sample system of the two regions of urban traffic. The adaptive concept is employed to estimate the upper bound of uncertainties and external disturbances. Then, the finite-time stability proof is performed by choosing the proper Lyapunov candidate function. Finally, numerical simulation results are carried out for a case study in Simulink/MATLAB to evaluate the proposed method in this paper.

2 Mathematical Preliminaries

Lemma 1: For each value $a_1, a_2, \dots, a_n \in \Re$ and $0 < q < 2$ we have, $|a_1|^q + |a_2|^q + \dots + |a_n|^q \geq (a_1^2 + a_2^2 + \dots + a_n^2)^{\frac{q}{2}}$ [27, 28].

Lemma 2: In the nonlinear system $\dot{x} = f(x), f(0) = 0, x \in \Re^n$ with initial conditions $x(0) = x_0$, if the Lyapunov candidate function $V(x)$ is globally positive definite, radially unbounded and only at $x = 0$ is zero, and the time derivative of the Lyapunov candidate function is as $\dot{V}(x) \leq -\rho_1 V^{\rho_2}(x)$, where ρ_1 is a positive number and ρ_2 is a constant between zero and one; hence the variable x of the system from any initial conditions, it reaches zero in a finite time, and since then it remains exactly equal to zero, i.e. $\lim_{t \rightarrow T} x \rightarrow 0$

and the upper bound of the settling time T will be as $T \leq \frac{V^{1-\rho_2}(x_0)}{\rho_1(1-\rho_2)}$ [29, 30].

Definition 1: The $sign(a)$ function is defined as $sign(a) = \begin{cases} 1; & a > 0 \\ 0; & a = 0 \\ -1; & a < 0 \end{cases}$ [31].

Definition 2: The $sig(a)$ function is represented as $sig^b(a) = |a|^b sign(a)$ [32].

3 Problem Statement

The two regions traffic model in the urban network has been presented in [17] in the form of Eq. (1).

$$\begin{cases} \dot{n}_{11} = q_{11}(t) + u_{21}M_{21}(t) - M_{11}(t) \\ \dot{n}_{12} = q_{12}(t) - u_{12}M_{12}(t) \\ \dot{n}_{21} = q_{12}(t) - u_{21}M_{21}(t) \\ \dot{n}_{22} = q_{22}(t) + u_{12}M_{12}(t) - M_{22}(t) \end{cases} \quad (1)$$

where $n_{ij}(t)$, $i, j = (1, 2)$ are the number of vehicles in region i with destination region j and n_i are the total number of existing vehicles in region i , i.e. $n_i = \sum_{j=1}^2 n_{ij}q_{ij}(t)$ are exogenous demand of vehicles from region i to region j . $M_{ij}(t)$ is the internal and external flow of the trips, which is defined as $M_{ij} = \frac{n_{ij}(t)}{n_i(t)}G_i(n_i(t))$. $G_i(n_i(t))$ are the complete flow of internal trips, which are introduced as $G_i(n_i(t)) = a_i n_i^3(t) + b_i n_i^2(t) + c_i n_i(t)$. The control inputs $u_{12}(t)$, $u_{21}(t)$ are also on the border between the two regions and aim to control the urban traffic system variables in a desirable condition. Note that the estimation of uncertainties and external disturbances are used in these control inputs.

By defining the changing variables of Eq. (2), we have:

$$\begin{cases} z_1(t) = n_{11}(t) + n_{21}(t) \\ z_2(t) = n_{12}(t) \\ z_3(t) = n_{21}(t) \\ z_4(t) = n_{12}(t) + n_{22}(t) \end{cases} \quad (2)$$

In consequence,

$$\begin{cases} \dot{z}_1(t) = q_{11}(t) + q_{21}(t) - M_{11}^z(t) + d_1(t) \\ \dot{z}_2(t) = q_{12}(t) - u_{12}M_{12}^z(t) + d_2(t) \\ \dot{z}_3(t) = q_{12}(t) - u_{21}M_{21}^z(t) + d_3(t) \\ \dot{z}_4(t) = q_{22}(t) + q_{12}(t) - M_{22}^z(t) + d_4(t) \end{cases} \quad (3)$$

where $d_p(t)$, $p = (1, 2, 3, 4)$ is the model of uncertainties and external disturbances. An upper bound is considered for each of these uncertainties and external disturbances, and their estimation is used in the control input. In other words, we have, $|d_p| \leq h_p$ and $h_p \leq \hat{h}_p \leq h_p^*$, where \hat{h}_p is the estimation of the upper bound of the uncertainties

and external disturbances and h_p^* is the upper bound of the estimation. Also, $M_{ij}^z(t)$ is as follows:

$$\begin{cases} M_{11}^z(t) = \frac{z_1(t) - z_3(t)}{z_1(t) - z_3(t) + z_2(t)} G_1(z_1(t) - z_3(t) + z_2(t)) \\ M_{12}^z(t) = \frac{z_2(t)}{z_1(t) - z_3(t) + z_2(t)} G_1(z_1(t) - z_3(t) + z_2(t)) \\ M_{21}^z(t) = \frac{z_3(t)}{z_4(t) - z_2(t) + z_3(t)} G_2(z_4(t) - z_2(t) + z_3(t)) \\ M_{22}^z(t) = \frac{z_4(t) - z_2(t)}{z_4(t) - z_2(t) + z_3(t)} G_2(z_4(t) - z_2(t) + z_3(t)) \end{cases} \quad (4)$$

In this paper, the control goal is $n_{11}(t) = -(1 + k_2)n_{21}(t)$, $n_{22}(t) = -(1 + k_1)n_{12}(t)$ where $k_i < -1$. In the next section, the control inputs are designed so that the system fulfills its control goal in a finite time. Also, the upper bound of the uncertainties and external disturbances is estimated using adaptive laws and their estimations in the control inputs.

4 Controller Design

Theorem 1: by assuming the system in Eq. (3) with the described conditions for it and considering the sliding surfaces in Eq. (5) and the control inputs in Eq. (6), and the adaptive laws in Eq. (7), consequently, the control goal of this article is guaranteed in a limited time. Also, the effect of the uncertainties and external disturbances is eliminated by estimating the upper bound.

$$\begin{cases} s_1 = z_4(t) + k_1 z_2(t) \\ s_2 = z_1(t) + k_2 z_3(t) \end{cases} \quad (5)$$

and control inputs are as follows:

$$\begin{cases} u_{12}(t) = \left(\frac{1}{k_1 M_{12}(t)}\right) \begin{pmatrix} (k_1 + 1)q_{12}(t) + q_{22}(t) - M_{22}(t) \\ + \hat{h}_2 |k_1| sig^{\alpha_2}(s_1) + \hat{h}_4 sig^{\alpha_4}(s_1) \end{pmatrix} \\ u_{21}(t) = \left(\frac{1}{k_2 M_{21}(t)}\right) \begin{pmatrix} (k_2 + 1)q_{21}(t) + q_{11}(t) - M_{11}(t) \\ + \hat{h}_3 |k_2| sig^{\alpha_3}(s_2) + \hat{h}_1 sig^{\alpha_1}(s_2) \end{pmatrix} \end{cases} \quad (6)$$

where α_p is positive control parameters and smaller than one, and the adaptive laws are as follows:

$$\begin{cases} \dot{\hat{h}}_1 = r_1 |s_2|^{\alpha_1+1} \\ \dot{\hat{h}}_2 = r_2 |s_1|^{\alpha_2+1} \\ \dot{\hat{h}}_3 = r_3 |s_2|^{\alpha_3+1} \\ \dot{\hat{h}}_4 = r_4 |s_1|^{\alpha_4+1} \end{cases} ; \begin{cases} r_1 \leq 1 \\ r_2 \leq |k_1| \\ r_3 \leq |k_2| \\ r_4 \leq 1 \end{cases} \quad (7)$$

Proof: To prove that the system reaches the sliding surfaces $s_i = 0$ in a finite time, and consequently reaching to control goal, the candidate Lyapunov function is considered as $V(x) = \frac{1}{2}s_1^2 + \frac{1}{2}s_2^2 + \sum_{p=1}^4 \frac{1}{2}\tilde{h}_p^2$. By Differentiating the Lyapunov function, there comes:

$$\dot{V}(x) = s_1\dot{s}_1 + s_2\dot{s}_2 + \sum_{p=1}^4 \hat{h}_p\tilde{h}_p \quad (8)$$

By applying control inputs and adaptive laws, we have

$$\begin{aligned} \dot{V}(x) &\leq |s_1|h_4 + |k_1||s_1|h_2 + |s_2|h_1 + |k_2||s_2|h_3 - \hat{h}_3|k_2||s_2|^{\alpha_3+1} - \\ &\hat{h}_1|s_2|^{\alpha_1+1} - \hat{h}_2|k_1||s_1|^{\alpha_2+1} - \hat{h}_4|s_1|^{\alpha_4+1} + \sum_{p=1}^4 \hat{h}_p\tilde{h}_p \quad \rightarrow \quad \dot{V}(x) \leq -\Delta_1|s_1| - \\ &\Delta_2|s_2| - \sum_{p=1}^4 \Delta_p\tilde{h}_p \quad \rightarrow \quad \dot{V}(x) \leq -\Delta_{min}(|s_1| + |s_2| + \sum_{p=1}^4 \tilde{h}_p) \quad \rightarrow \quad \dot{V}(x) \leq \\ &-\Delta_{min}(|s_1|^2 + |s_2|^2 + \sum_{p=1}^4 \tilde{h}_p^2)^{\frac{1}{2}} \quad \rightarrow \quad \dot{V}(x) \leq -\Delta_{min}(2V(x))^{\frac{1}{2}} \end{aligned} \quad (9)$$

where $\Delta_{min} = \min(\Delta_1, \Delta_2, \dots, \Delta_6)$. As a result, it is guaranteed by Lemma 1 that the upper bound of the uncertainties is estimated in the finite time as well as the system reaches the sliding surface $s_i = 0$ in a finite time. ■

5 Simulations

To perform numerical simulation, the Simulink environment of the MATLAB software has been used by considering the numerical solution ode4 and step 0.1 for 4000 s. The uncertainties and external disturbances model and external disturbances have been selected as $d_p = 0.01\cos(t)$, and the control parameters as $r_1 = r_2 = r_3 = r_4 = 0.005$, $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \frac{11}{1003}$, $k_1 = -1.25$, $k_2 = -1.5$. Figure 1 shows travel demands, Fig. 2 presents the number of existing vehicles in each region, and Fig. 3 displays the designed control inputs.

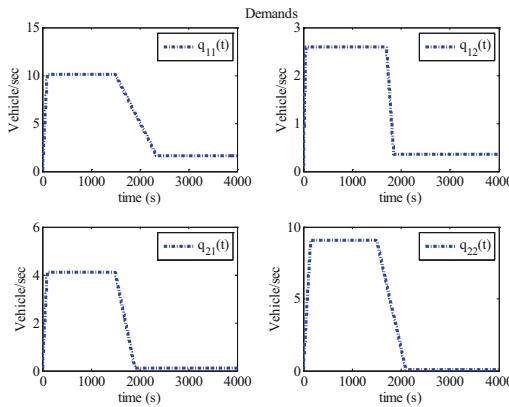


Fig. 1. The travel demands.

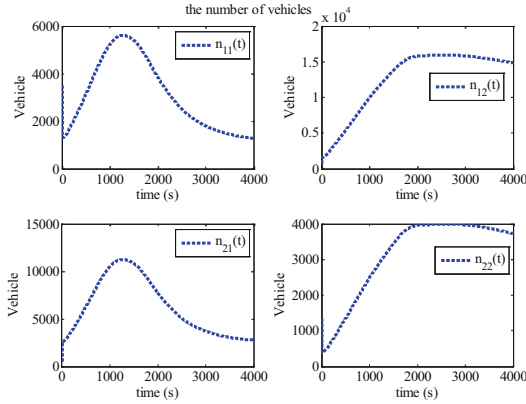


Fig. 2. The number of vehicles in each region.

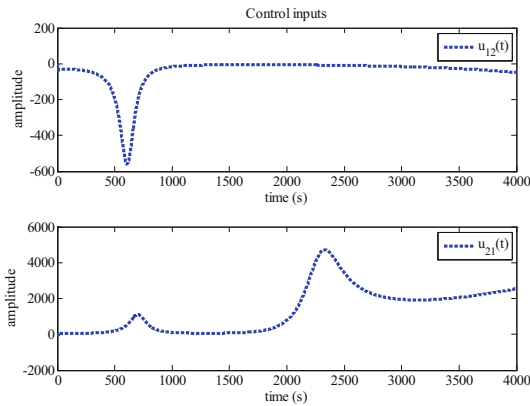


Fig. 3. The control inputs.

6 Discussion

The amount of considered travel demand changes naturally over time (see Fig. 1). Indeed, travel demand is low initially, then gradually increases as time passes and reaches its peak. It follows by a decrease gradually afterward. By considering this form of travel demand, the number of vehicles concerning time in different regions is shown in Fig. 2. As shown in Fig. 2, the number of vehicles increases and gradually decreases as time passes. This is a natural thing that practically happens for urban traffic networks. This issue by itself confirms the correctness of the article design. The issue of diversity in management and control of the different regions is critical because traffic control measures change with the change of traffic time and flow, like traffic light cycle, etc. As shown in Fig. 3, control inputs have changed based on time and amplitude. Figure 3 shows that the amplitude of regions 1 and 2 concerning time and considering the considered parameters in Eq. (6) are different from each other and change over time. In such a way, the amplitude of region 1 decreases as time passes until about 600 s and then sharply increases. Also, the

amplitude of region 2 increases from approximately 2,000 s to 2,300 s and then decreases by 2600 s. Subsequently, its amplitude remains almost constant. In this paper, the control inputs and adaptive laws are designed by incorporating the adaptive concept and TSMC scheme, which lead to estimating the upper bounds of the uncertainties and external disturbances of the two-region system and controlling the traffic flow in a finite time. In this design, the proposed control inputs applicable to implementation can respond to different travel demands in urban traffic flows.

7 Conclusion

In this paper, theoretically, a Terminal Sliding Mode Control method is designed to control traffic flow in urban networks. The proposed control method controlled the model very well. The simulation results show that this controller is feasible for controlling the traffic flow. The Lyapunov function has been employed for stability-proofing the controller. For future work, it is recommended to focus on optimizing control inputs or specific cost functions.

References

1. Fusco, G., Colombaroni, C., Isaenko, N.: Short-term speed predictions exploiting big data on large urban road networks. *Transport. Res. Part C: Emerg. Technol.* **73**, 183–201 (2016)
2. He, F., Yan, X., Liu, Y., Ma, L.: A traffic congestion assessment method for urban road networks based on speed performance index. *Procedia Eng.* **137**, 425–433 (2016)
3. Jovanović, A., Nikolić, M., Teodorović, D.: Area-wide urban traffic control: a Bee Colony Optimization approach. *Transport. Res. Part C: Emerg. Technol.* **77**, 329–350 (2017)
4. Ben-Akivai, M., Bowman, J.L., Gopinath, D.: Travel demand model system for the information era. *Transportation* **23**(3), 241–266 (1996)
5. Ferguson, E.: Transportation demand management planning, development, and implementation. *J. Am. Plann. Assoc.* **56**(4), 442–456 (1990)
6. Thonhofer, E., Palau, T., Kuhn, A., Jakubek, S., Kozek, M.: Macroscopic traffic model for large scale urban traffic network design. *Simul. Model. Pract. Theory* **80**, 32–49 (2018)
7. Gayah, V.V., Gao, X.S., Nagle, A.S.: On the impacts of locally adaptive signal control on urban network stability and the macroscopic fundamental diagram. *Transport. Res. Part B: Methodol.* **70**, 255–268 (2014)
8. Mitsakis, E., Salanova, J.M., Giannopoulos, G.: Combined dynamic traffic assignment and urban traffic control models. *Procedia Soc. Behav. Sci.* **20**, 427–436 (2011)
9. Zhong, R., Chen, C., Huang, Y., Sumalee, A., Lam, W., Xu, D.: Robust perimeter control for two urban regions with macroscopic fundamental diagrams: a control-Lyapunov function approach. *Transport. Res. Part B: Methodol.* **117**, 687–707 (2018)
10. Castro, G.B., Hirakawa, A.R., Martini, J.S.: Adaptive traffic signal control based on bio-neural network. *Procedia Comput. Sci.* **109**, 1182–1187 (2017)
11. Lin, S., Xi, Y.: An efficient model for urban traffic network control. *IFAC Proceedings Volumes* **41**(2), 14066–14071 (2008)
12. Tettamanti, T., Varga, I., Péni, T., Luspay, T., Kulcsar, B.: Uncertainty modeling and robust control in urban traffic. *IFAC Proceed.* Vol. **44**(1), 14910–14915 (2011)
13. Zhou, Z., Lin, S., Li, D., Xi, Y.: A congestion eliminating control method for large-scale urban traffic networks. *IFAC Proceed.* Vol. **46**(13), 496–501 (2013)

14. Zhou, X., Ye, B.-L., Lu, Y., Xiong, R.: A novel MPC with chance constraints for signal splits control in urban traffic network. *IFAC Proceedings Volumes* **47**(3), 11311–11317 (2014)
15. Chai, H., Zhang, H.M., Ghosal, D., Chuah, C.-N.: Dynamic traffic routing in a network with adaptive signal control. *Transport. Res. Part C: Emerg. Technol.* **85**, 64–85 (2017)
16. Tian, R., Zhang, X.: Design and evaluation of an adaptive traffic signal control system—a case study in hefei, china. *Transport. Res. Procedia* **21**, 141–153 (2017)
17. Aalipour, A., Kebriaei, H., Ramezani, M.: Nonlinear robust traffic flow control in urban networks. *IFAC-PapersOnLine* **50**(1), 8537–8542 (2017)
18. Chen, X., Yu, H., Hao, F.: Prescribed-time event-triggered bipartite consensus of multiagent systems. *IEEE Transactions on Cybernetics* (2020)
19. Wang, Y., Song, Y., Hill, D.J., Krstic, M.: Prescribed-time consensus and containment control of networked multiagent systems. *IEEE Trans. Cybern.* **49**(4), 1138–1147 (2018)
20. F. Amato, R. Ambrosino, M. Ariola, C. Cosentino, and G. De Tommasi, *Finite-time stability and control*. Springer, 2014
21. Muralidharan, A., Pedarsani, R., Varaiya, P.: Analysis of fixed-time control. *Transport. Res. Part B: Methodol.* **73**, 81–90 (2015)
22. Abadi, A.S.S., Hosseinabadi, P.A., Mekhilef, S.: Fuzzy adaptive fixed-time sliding mode control with state observer for a class of high-order mismatched uncertain systems. *Int. J. Control Autom. Syst.* **18**, 2492–2508 (2020)
23. Hosseinabadi, P.A., Pota, H., Mekhilef, S., Schwartz, H.: Fixed-time observer-based control of DFIG-based wind energy conversion systems for maximum power extraction. *Int. J. Electr. Power Energy Syst.* **146**, 108741 (2023)
24. Alinaghi Hosseinabadi, P., Soltani Sharif Abadi, A., Schwartz, H., Pota, H., Mekhilef, S.: Fixed-time sliding mode observer-based controller for a class of uncertain nonlinear double integrator systems. *Asian J. Control* **25**, 3052 (2023)
25. Sánchez-Torres, J.D., Sanchez, E.N., Loukianov, A.G.: Predefined-time stability of dynamical systems with sliding modes. In: 2015 American control conference (ACC), pp. 5842–5846: IEEE (2015)
26. Abadi, A.S.S., Hosseinabadi, P.A., Mekhilef, S., Ordys, A.: A new strongly predefined time sliding mode controller for a class of cascade high-order nonlinear systems. *Archives of Control Sciences*, pp. 599–620 (2020)
27. Bhat, S.P., Bernstein, D.S.: Continuous finite-time stabilization of the translational and rotational double integrators. *IEEE Trans. Autom. Control* **43**(5), 678–682 (1998)
28. Abadi, A.S.S., Mehrizi, M.H., Hosseinabadi, P.A.: Fuzzy adaptive terminal sliding mode control of SIMO nonlinear systems with TS fuzzy model. In: 2018 6th Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS), pp. 185–189. IEEE (2018)
29. Qiao, L., Zhang, W.: Adaptive non-singular integral terminal sliding mode tracking control for autonomous underwater vehicles. *IET Control Theory Appl.* **11**(8), 1293–1306 (2017)
30. Abadi, A.S.S., Hosseinabadi, P.A.: Fuzzy adaptive finite time control ship fin stabilizing systems model of fuzzy Takagi-Sugeno with unknowns and disturbances. In: 2018 6th Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS), pp. 33–36. IEEE (2018)
31. Abadi, A.S.S., PHosseinabadi, P.A., Mekhilef, S.: Two novel AOTSMC of photovoltaic system using VSC model in smart grid. In: 2017 Smart Grid Conference (SGC), pp. 1–6. IEEE (2017)
32. Alinaghi Hosseinabadi, P., Soltani Sharif Abadi, A., Mekhilef, S., Pota, H.R.: Fixed-time adaptive robust synchronization with a state observer of chaotic support structures for offshore wind turbines. *J. Control Autom. Electr. Syst.* **32**(4), 942–955 (2021)



An Application of the Dynamic Decoupling Techniques for a Nonlinear TITO Plant

Szymon Król^(✉) and Paweł Dworak^(ID)

West Pomeranian University of Technology in Szczecin, ul. 26 Kwietnia 10, Szczecin, Poland

{ks46884,pawel.dworak}@zut.edu.pl

Abstract. In this paper the basic dynamic decoupling methods of two-inputs two-outputs TITO dynamic systems were presented. The article describes the problems of implementing the decoupling techniques for nonlinear thermal plants and shows the solution for the air heater. The new identification method based on the gain-scheduling strategy was proposed further on.

Keywords: dynamic decoupling · TITO plants · nonlinear systems · air heater · gain scheduling

1 Introduction

The synthesis of a closed-loop control system for the multiple-input multiple-output MIMO dynamic plant may be a difficult task, when one of the inputs affects more than just one of the outputs. Therefore, to achieve the desired system's behaviour, the decoupling techniques are widely used. The dynamic decoupling methods include both steady and transition states of a system. Despite the kind, the decoupling usually splits a MIMO plant into a group of a single-input single-output SISO systems, that may be controlled individually or into a group of a smaller MIMO plants, when the main system is non-square. [4] presents a variety of the dynamic decoupling methods, from which a so-called autonomization can be distinguished. It is the most demanding and desirable case, in which an influence only between one input and one output exists, i.e. a group of SISO plants is created.

Some particular group among the MIMO plants are two-input two-outputs TITO systems, whose structure often allows for easier use of a different decoupling algorithms. In [7,9,11,13–17] three basic methods are mentioned: ideal, simplified and inverted. Each method applies to a plant model in the form of a transfer function matrix. The resulting dynamic system, combined of the dynamic decoupler and a plant, has a diagonal transfer function matrix form.

However, the cited methods refers to linear systems, which means, that a single decoupler will work throughout the whole state space. The decoupled process may be described with the linear state equations, as shown in [3].

In case of nonlinear plants their parameters change as a function of the operating point. Moreover, a perfect decoupling of a nonlinear plant may be impossible due to a strong non-linearity and a high coupling level, therefore the dynamic decoupling of a nonlinear system usually does not completely cancel out the cross-couplings interferences, but reduces them. Nevertheless, this reduction is often sufficient and satisfies the requirements.

A nonlinear plant can be linearized globally with the nonlinear state-feedback [10, 18] or in the certain operating point with the Taylor series [12]. The problem is, that both methods require a nonlinear state equations, that can be difficult to derive in case of a complex processes. One possible way to obtain the linear model of the nonlinear plant is the experimental modeling [6, 7, 15]. The model is simplified to the n -th order two-dimensional inertial system, which also simplifies the calculations of the decoupler system, because the transfer function matrix is already given. The more experimental data is gathered, the better the model fits the plant. A sufficient amount of the described models may cover the needed state space part, but for every operating point a new decoupler has to be established. This means, that the decoupler system must be switched every time the plant changes its state, which brings the switching algorithm to the decoupling process and creates a switchable decoupler structure. Such a structure can be based on the neural network as well as on the fuzzy logic, which may be the topic of the further researches.

The article presents the problems of implementing decoupling systems for non-linear TITO objects. On the example of an air heater, the method of identifying the parameters of its model and synthesis of the gain-scheduling adaptive decoupling precompensator were presented. The work ends with the presentation of the results of the experimental verification of the developed method and the final conclusions.

2 Basic Dynamic Decoupling Techniques

A typical TITO dynamic plant is usually described as a transfer function matrix:

$$\mathbf{G}(s) = \begin{bmatrix} G_{11}(s) & G_{12}(s) \\ G_{21}(s) & G_{22}(s) \end{bmatrix} \quad (1)$$

where each transfer function $G_{ij}(s)$ is a n -th order inertia with the dead-time:

$$G_{ij}(s) = \frac{k_{ij}}{(s + T_{1ij})(s + T_{2ij}) + \dots + (s + T_{nij})} e^{-s\tau_{ij}} \quad (2)$$

$$i, j = 1, 2 \wedge n \in \mathbf{N}_+ \wedge T_{nij}, \tau_{ij} \in \mathbf{R}_+$$

as mentioned in [1, 5–7, 9, 11, 13–17]. Presented transfer function can be a result of the suggested experimental modeling [6, 7, 15] or be derived directly from state space model or differential equations in case of a linear plants [3, 12]. Note, that

Eq. (2) shows the general form of a transfer function and the most common form (and used further in this paper) is FOPDT, where $n = 1$. The block diagram of the typical TITO plant is presented in Fig. 1.

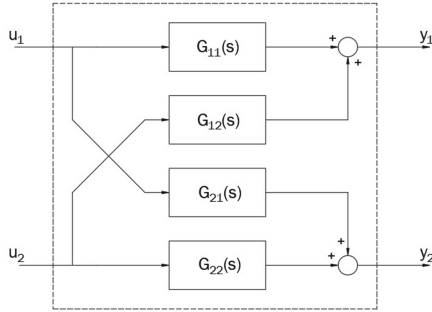


Fig. 1. Block diagram of the TITO system

The idea behind the dynamic decoupling regarding to a TITO dynamic plants is to calculate a decoupler matrix $\mathbf{D}(s) \in \mathbf{C}^{2 \times 2}$, such that:

$$\mathbf{H}(s) = \mathbf{G}(s)\mathbf{D}(s) = \begin{bmatrix} H_{11}(s) & 0 \\ 0 & H_{22}(s) \end{bmatrix} \tag{3}$$

so the decoupled system can be treated as two independent SISO systems. By rewriting the Eq. (3), the decoupler matrix is given as:

$$\mathbf{D}(s) = \mathbf{G}^{-1}(s)\mathbf{H}(s) = \frac{1}{|\mathbf{G}(s)|} \begin{bmatrix} G_{22}(s)H_{11}(s) & -G_{12}(s)H_{22}(s) \\ -G_{21}(s)H_{11}(s) & G_{11}(s)H_{22}(s) \end{bmatrix} \tag{4}$$

Assuming, that the decoupled system’s dynamics is the exact plant’s main diagonal dynamics, i.e.:

$$\begin{aligned} H_{11}(s) &= G_{11}(s) \\ H_{22}(s) &= G_{22}(s) \end{aligned} \tag{5}$$

the ideal decoupler formula is established:

$$\mathbf{D}(s) = \begin{bmatrix} \frac{G_{11}(s)G_{22}(s)}{G_{11}(s)G_{22}(s)-G_{12}(s)G_{21}(s)} & \frac{-G_{12}(s)G_{22}(s)}{G_{11}(s)G_{22}(s)-G_{12}(s)G_{21}(s)} \\ \frac{-G_{21}(s)G_{22}(s)}{G_{11}(s)G_{22}(s)-G_{12}(s)G_{21}(s)} & \frac{G_{11}(s)G_{22}(s)}{G_{11}(s)G_{22}(s)-G_{12}(s)G_{21}(s)} \end{bmatrix} \tag{6}$$

The ideal decoupler’s block diagram is presented in Fig. 2 a).

The decoupled system’s dynamics may be however different. The goal is in fact to minimize the cross-couplings impact, but the resulting dynamics of

the decoupled system does not matter from a closed-loop control perspective. Therefore, it is possible to assume, that:

$$D_{11}(s) = D_{22}(s) = 1 \tag{7}$$

which, combined with Eq. (4), derives the simplified decoupler formula:

$$\mathbf{D}(s) = \begin{bmatrix} 1 & -\frac{G_{12}(s)}{G_{11}(s)} \\ -\frac{G_{21}(s)}{G_{22}(s)} & 1 \end{bmatrix} \tag{8}$$

and the decoupled system is given as:

$$\mathbf{H}(s) = \begin{bmatrix} G_{11}(s) - \frac{G_{12}(s)G_{21}(s)}{G_{22}(s)} & 0 \\ 0 & G_{22}(s) - \frac{G_{12}(s)G_{21}(s)}{G_{11}(s)} \end{bmatrix} \tag{9}$$

The simplified decoupler’s block diagram is presented in Fig. 2 b).

Last known decoupling technique is the inverted decoupling. This method combines the simplicity of the simplified decoupler with the performance of the ideal decoupler. The inverted decoupler is created with the simplified decoupler from Eq. (8) by simply reversing it, which forms a decoupled system with main diagonal elements from Eq. (5). The inverted decoupler’s block diagram is presented in Fig. 2 c). All described decoupling techniques are presented in [1, 7, 9, 11, 13–17].

Regardless of the chosen decoupling method, defining the τ_{ij} value may cause a problem due to a high complexity of the ideal decoupler equations and the possibility of the appearance of the positive values in the simplified method. The solution was proposed in [13] by adopting a function:

$$v(L) = \begin{cases} L & \text{if } L > 0 \\ 0 & \text{if } L \leq 0 \end{cases} \tag{10}$$

Using the function (10), the decoupler matrix in both methods can be transformed into:

$$\mathbf{D}(s) = \begin{bmatrix} \frac{N_{11}(s)}{M_{11}(s)} e^{-v(\tau_{22}-\tau_{21})s} & \frac{N_{12}(s)}{M_{12}(s)} e^{-v(\tau_{12}-\tau_{11})s} \\ \frac{N_{21}(s)}{M_{21}(s)} e^{-v(\tau_{21}-\tau_{22})s} & \frac{N_{22}(s)}{M_{22}(s)} e^{-v(\tau_{11}-\tau_{21})s} \end{bmatrix} \tag{11}$$

The simplified decoupling method always provides the stability of the decoupler. [5] introduces a coupling level factor ϵ . Depending on the ϵ value, the coupling may be weak, intermediate or strong. The ϵ factor was described as a gain value of the plant’s antidiagonal transfer functions:

$$\mathbf{G}(s) = \begin{bmatrix} \frac{k_{11}}{T_{11}s+1} e^{-s\tau_{11}} & \frac{\epsilon}{T_{12}s+1} e^{-s\tau_{12}} \\ \frac{\epsilon}{T_{21}s+1} e^{-s\tau_{21}} & \frac{k_{22}}{T_{22}s+1} e^{-s\tau_{22}} \end{bmatrix} \tag{12}$$

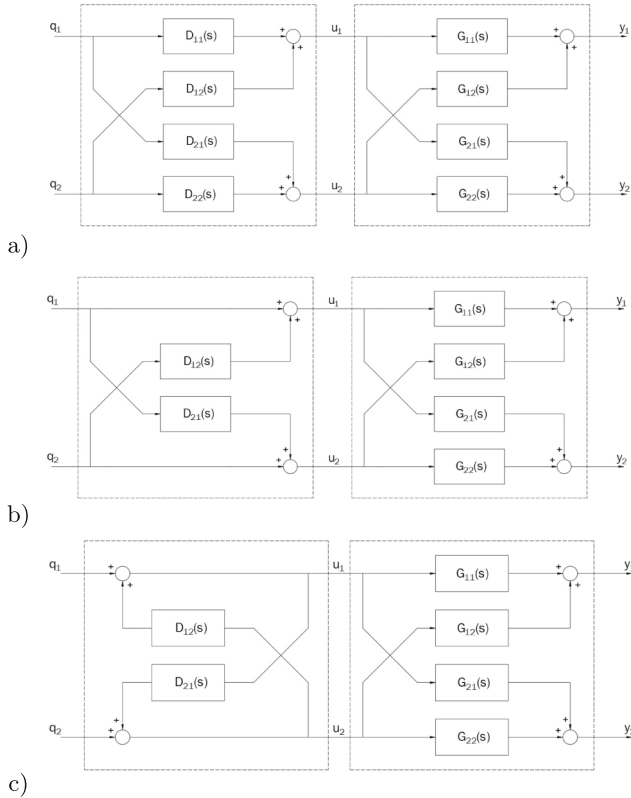


Fig. 2. Block diagram of the a) ideal b) simplified c) inverted decoupler and plant

For such a plant, the simplified decoupler’s antidiagonal characteristic polynomials are given as:

$$\begin{aligned}
 M_{12}(s) &= s + \frac{1}{T_{12}} \\
 M_{21}(s) &= s + \frac{1}{T_{21}}
 \end{aligned}
 \tag{13}$$

which means, that the decoupler’s stability does not depend on the coupling level. It is always stable, because T_{12}, T_{21} are expected to be greater than zero in Eq. (2). The $M_{12}(s), M_{21}(s)$ roots (the decoupler’s poles) are located in the left half of the complex plane.

In the ideal decoupling technique both $D_{12}(s)$ and $D_{21}(s)$ have the same characteristic polynomial:

$$M(s) = s^2 + \frac{T_{12} + T_{21} - \epsilon^2(T_{11} + T_{22})}{T_{21}T_{12} - \epsilon^2T_{11}T_{22}}s + \frac{1 - \epsilon^2}{T_{21}T_{12} - \epsilon^2T_{11}T_{22}}
 \tag{14}$$

Now the factor ϵ affects the location of the decoupler poles. The Eq. (14) can be rewritten as:

$$M(s) = (s - p_1)(s - p_2) \quad (15)$$

The ϵ value is bounded to the $[0, 1]$ interval, so determining the following limits is possible:

$$\begin{aligned} \lim_{\epsilon \rightarrow 0} p_1 &< 0 \\ \lim_{\epsilon \rightarrow 0} p_2 &< 0 \\ \lim_{\epsilon \rightarrow 1} p_1 &< 0 \\ \lim_{\epsilon \rightarrow 1} p_2 &> 0 \end{aligned} \quad (16)$$

which proves, that the ideal decoupler may be unstable, when the plant is highly coupled. It is an unwanted case, which can lead to the instability of the decoupled system.

Besides the stability, the $G_{12}(s)$ and $G_{21}(s)$ gains will usually differ. Therefore, the ϵ definition should be extended as follows:

$$\begin{aligned} \epsilon_1 &= \frac{k_{12}}{k_{11}} \\ \epsilon_2 &= \frac{k_{21}}{k_{22}} \end{aligned} \quad (17)$$

so that the coupling factors are the ratios of the cross-coupling's and disturbed system's gains. Calculating the ϵ value may be an indicator, which decoupling method should be chosen. For instance, the unstable decoupled system can be stabilized with the proper closed-loop control, so the instability is not an issue in this case. [8] shows an example, where a strongly nonlinear unstable system is decoupled and stabilized with an adaptive controller. However, in open-loop control a decoupled system must be stable, especially when controlled by the operator.

3 Synthesis of the Dynamic Decoupler for the Air Heater System

The nonlinearity of a thermal processes consists not only in the changes of the parameters in a function of the operating point, but also in the actual state of heating or cooling. Therefore, the classical first order inertia with dead-time FOPDT model can be extended to a form, which includes actual input signal changes [19]. The extended model may result in a high complexity of the switchable decoupler structure. The experimental modeling of the heating state only may be a way to obtain a sufficient linear model. Further part of this paper presents an example of the thermal plant - the air heater and explains the need of creating the switchable decoupling structure. Moreover, the gain-scheduling identification method was introduced.

3.1 Structure of the Air Heater System

The presented plant has two inputs: first to control a fan, second to control the heating coil. The two outputs are measured with the PC-50 pressure converter and the PT-100 temperature resistance sensor. The exact diagram of the discussed system is presented in Fig. 3. The system uses the Advantech’s PCI-1710HG DAQ card and the Schneider Electric’s Zelio Analog Interface Modules. I/O signals are generated and gathered using Matlab and Simulink environment. All signals are normalized to the $(4 \div 20)$ mA interval. However, the currents are converted to the voltage signals in $(0 \div 10)$ V interval and normalized to the $(0 \div 1)$ dimensionless quantity interval in Simulink.

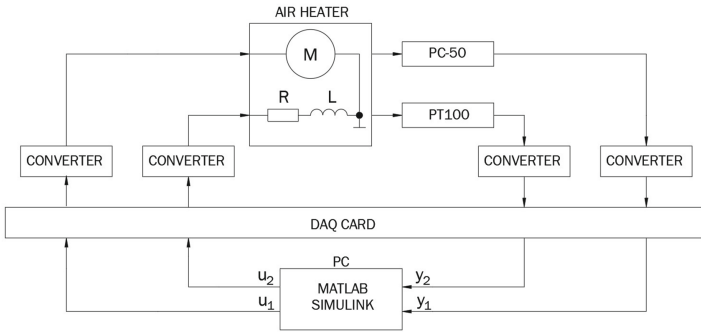


Fig. 3. Block diagram of the air heater system

In the further part of this paper $u_1(t)$ is the fan control signal (1st input), $u_2(t)$ is a heating coil control signal (2nd input), $y_1(t)$ is the air pressure (1st output) and $y_2(t)$ is the air temperature (2nd output). The temperature changes do not interfere with the pressure level and the pressure influence is negative, therefore the Eq. (1) simplifies:

$$\mathbf{G}(s) = \begin{bmatrix} G_{11}(s) & 0 \\ -G_{21}(s) & G_{22}(s) \end{bmatrix} \tag{18}$$

Due to system’s model simplification, both ideal and simplified decoupler equations converge:

$$\mathbf{D}(s) = \begin{bmatrix} 1 & 0 \\ \frac{G_{21}(s)}{G_{22}(s)} & 1 \end{bmatrix} \tag{19}$$

It is a special case, where the choice of the decoupling method does not matter. Moreover, both methods provide the decoupler’s stability due to Eq. (13). The block diagram of the air heater with the decoupler is shown in Fig. 4.

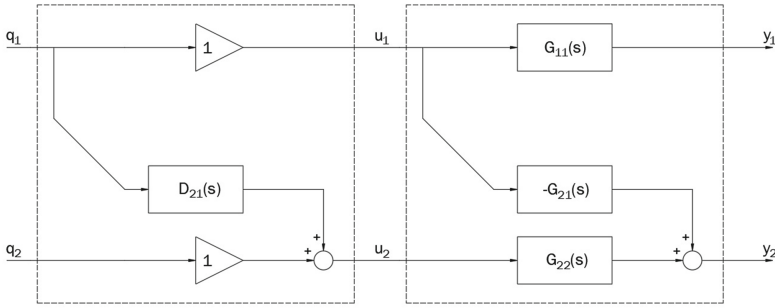


Fig. 4. Block diagram of the air heater model and decoupler

3.2 Identification Procedure

Each element of the Eq. (18) matrix was identified as FOPDT (Eq. (2) for $n = 1$). The modeling process was based on the method described in [2] and consists of a few assumptions. The main goal was to create the quickest and the simplest identification method. First of all, it is assumed that $\tau_{21} \leq \tau_{22}$, so the function $v(L)$ from Eq. (10) always results in zero for $D_{21}(s)$. Based on the observations combined with the plant knowledge, and mostly to simplify the identification process, a linear relationship between T_{21} and T_{22} could be assumed:

$$T_{21} = \alpha T_{22}, \alpha = 0.95 \tag{20}$$

The identification was divided into following steps:

1. Set the fan control signal $u_1(t)$ to zero. Increase the heating coil control signal $u_2(t)$ by a small value Δu_2 each time the temperature enters the steady state. Continue increasing $u_2(t)$ to the moment the temperature achieves the measurement boundary. Each Δu_2 gives a new $G_{22}(s)$ model. It is assumed, that the obtained models have the exact $H_{22}(s)$ decoupled system's dynamics. Due to Eq. (18) and (19) $H_{11}(s) = G_{11}(s)$.

2. Divide the $u_1(t)$ signal range into n intervals:

$$u_1(t) = k(t), k \in \{\Delta u_1, 2\Delta u_1, \dots, n\Delta u_1\} \tag{21}$$

Repeat the experiment from Step 1. for the non-zero $u_1(t)$ signals.

3. Compare the temperature steady-states values from Step 2. with the reference values from Step 1., so the cross-coupling gain can be established:

$$k_{21} = \frac{\Delta y_2}{\Delta u_1} \tag{22}$$

The idea of the described method is presented in Fig. 5.

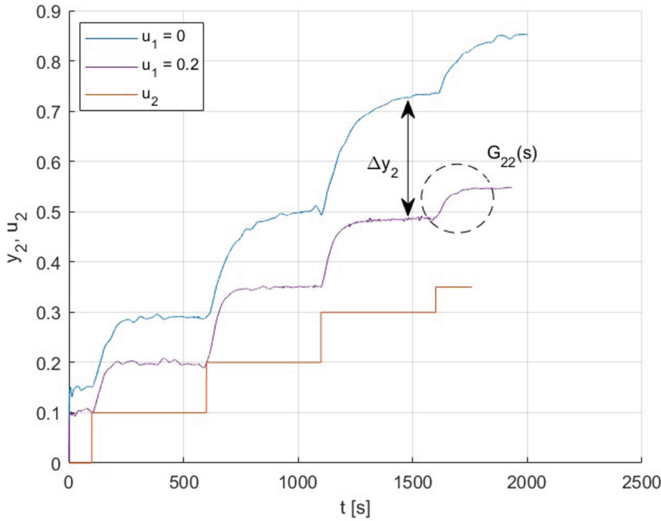


Fig. 5. Identification idea

By applying the described method the 32 linear models of the air heater were obtained. The gathered identification data allows for the evaluation of the dynamic decouplers for the 32 operating points, which is a sufficient amount to successfully decouple the nonlinear thermal plant.

3.3 Switchable Adaptive Decoupler

The switchable decoupler structure can be implemented in a variety of ways. It may be based on simple inputs values comparison and switching, neural networks or fuzzy logic. As the operating point changes, the decoupler has to adapt to the changing plant's parameters. For the air heater, from Eq. (19) and (20):

$$D_{21}(s) = \frac{G_{21}(s)}{G_{22}(s)} = \frac{k_{21}T_{22}s + k_{21}}{0.95k_{22}T_{22}s + k_{22}} \tag{23}$$

the $k_{21}, k_{22}, T_{21}, T_{22}$ values must change, as they are functions of $(q_1(t), q_2(t))$, i.e.: $k_{21} = f_1(q_1(t), q_2(t)), T_{21} = f_2(q_1(t), q_2(t))$ etc.

It was assumed, that the decoupler operates in the open-loop and does not take the output signals values into account, i.e. the system's input signals values define the current operating point. The adaptive decoupler structure was created using Mamdani fuzzy logic systems, where each fuzzy system changes a single decoupler's coefficient value, depending on the current input values. The input fuzzy sets correspond to the input signals values from the identification process, while the output fuzzy sets correspond to the established coefficients values. The fuzzy rules set simply describes the identification process and binds the input signals values with the corresponding coefficients values. However, the designed

fuzzy decoupler is not the subject of this work and will not be further presented in details.

4 Experimental Results

In the further experiment the decoupled system with switchable fuzzy logic based decoupler structure (FD), the decoupled system with single decoupler (SD) and the coupled system (C) were compared. During the experiments the $q_2(t)$ signal remained constant, but the $q_1(t)$ was varying over and beyond the chosen operating point. The results are presented in Fig. 6, where $q_{1,2}(t)$ are decoupler's inputs, $u_{1,2}(t)$ are plant's inputs and $y_{1,2}(t)$ are plant's outputs.

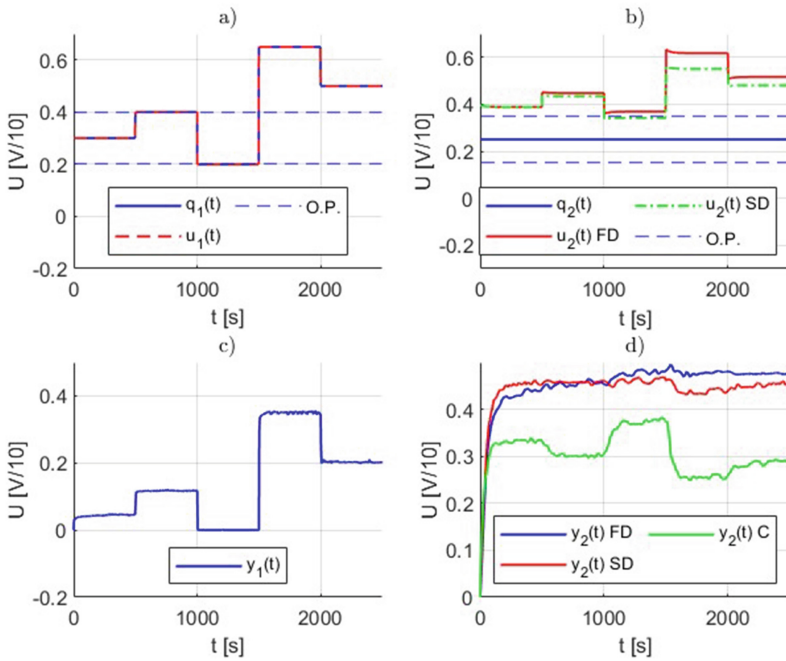


Fig. 6. Comparison of the coupled and decoupled air heater system

Both decoupler's structures (FD, SD) give better results, i.e. smaller temperature deviations compared to the coupled plant (Fig. 6 d)). The differences in systems' behaviours are most noticeable when $q_1(t)$ changes around 500th, 1000th, 1500th and 2000th second (Fig. 6 a)). The biggest change of the $q_1(t)$ value in 1500th second causes the change of the current operating point and further the change of the plant's and decoupler's parameters. The fuzzy decoupler (FD) took it into account, hence the $y_2(t)$ deviations after 1500th second are significantly smaller. The mentioned non-linearity of the thermal plant can be

observed, as the switchable structure adopted to the new operating point. Also the need of creating such a structure is explained.

5 Conclusion

The dynamic decoupling of nonlinear plants may cause a need of implementing an experimental modeling method, when the nonlinear state model is unknown or too complex to derive. The identification process of a thermal plant can be simplified using the proposed gain-scheduling based method, which is very easy to implement on today's hardware, for example on a PLC controller. It is also very convenient when the system's outputs and inputs are bounded. The described method omits the heating and cooling states of a thermal plant and proves, that identifying heating states only is sufficient. Moreover, this approach simplifies further creation of the switchable decoupler structure. Such a method may be a basis for a future researches on the self-decoupling algorithms.

References

1. Alawad, N., Alseady, A.: Fuzzy controller of model reduction distillation column with minimal rules. *Appl. Comput. Sci.* **16**(2), 80–94 (2020). <https://doi.org/10.23743/acs-2020-14>
2. Álvarez de Miguel, S., Mollocana Lara, J.G., García Cena, C.E., Romero, M., García de María, J.M., González-Aguilar, J.: Identification model and PI and PID controller design for a novel electric air heater. *Automatika* **58**(1), 55–68 (2017)
3. Amézquita-Brooks, L.A., Ugalde-Loo, C.E., Licéaga-Castro, E., Licéaga-Castro, J.: In-depth cross-coupling analysis in high-performance induction motor control. *J. Franklin Inst.* **355** (2018)
4. Bańka, S.: *Sterowanie wielowymiarowymi układami dynamicznymi. Ujęcie wielomianowe*, Wydawnictwo Uczelniane Politechniki Szczecińskiej (2007)
5. Berner, J., Soltész, K., Hägglund, T., Åström, K.J.: Autotuner identification of TITO systems using a single relay feedback experiment, 2405-8963 IFAC (2017)
6. Chen, M., Cui, S., Duan, H., Liu, J., Liu, Y.: Study on decoupling control system of temperature and pH concentration in photoreactive biological apparatus. In: *IEEE 5th Information Technology and Mechatronics Engineering Conference* (2020)
7. Chen, Q., Wang, Y., Xu, Z., Chen, X., Zhang, T., Zhang, Y.: Modeling, simulation and decoupling control of resistance furnace using Matlab and Simulink. In: *6th International Conference on Automation, Control and Robotics Engineering* (2021)
8. Dworak, P.: About Dynamic Decoupling of a Nonlinear MIMO Dynamic Plant, 978-1-4799-5081-2/14 IEEE (2014)
9. Effective Use of MPC for Dynamic Decoupling of MIMO Systems. *Elektronika IR Elektrotechnika* **25**(2) (2019). ISSN 1392-1215
10. Ermeydan A., Kaba A.: Feedback linearization of a quadrotor. In: *5th International Symposium on Multidisciplinary Studies and Innovative Technologies* (2021)
11. Hariz, M.B., Bouani, F.: Design of controllers for decoupled TITO systems using different decoupling techniques, 978-1-4799-8701-6/15 IEEE (2015)
12. Kaczorek, T., Dzieliński, A., Dąbrowski, W., Łopatka, R.: *Podstawy teorii sterowania*, Wydanie drugie zmienione Wydawnictwa Naukowo-Techniczne Warszawa (2005)

13. Li Z., Chen Y.: Ideal, simplified and inverted decoupling of fractional order TITO processes. In: Proceedings of the 19th World Congress The International Federation of Automatic Control (2014)
14. Lokesh, S.K., Sharma, S., Padhy, P.K.: Study of Different Decoupling Techniques for TITO Time-delay System, International Conference on Control, Automation, Power and Signal Processing (2021)
15. Mahapatro S. R., Subudhi B., Ghosh S., Dworak P.: A Comparative Study of Two Decoupling Control Strategies for a Coupled Tank System, 978-1-5090-2597-8/16 IEEE (2016)
16. Noeding, M., Martensen, J., Lemnke, N., Tegethoff, W., Koehler, J.: Selection of decoupling control methods suited for automated design for uncertain TITO processes. In: IEEE 14th International Conference on Control and Automation (2018)
17. Sharma A., Padhy P. K.: Design and implementation of PID controller for the decoupled two input two output control process, 978-1-5090-4426-9/17 IEEE (2017)
18. Shu-qing, L., Sheng-xiu, Z.: A simplified state feedback method for nonlinear control based on exact feedback linearization. In: International Conference on Computer Application and System Modeling (2010)
19. Skoczowski, S.: Dwustawna regulacja temperatury. Wydawnictwa naukowo-techniczne, Warszawa (1977)



Attitude Control of an Earth Observation Satellite with a Solar Panel

Zbigniew Emirsajłow¹, Tomasz Barciński², and Nikola Bukowiecka³

¹ West Pomeranian University of Technology, Szczecin, Poland
zbigniew.emirsajlow@zut.edu.pl

² Space Research Center, Polish Academy of Sciences, Warsaw, Poland
tbarcinski@cbk.waw.pl

³ West Pomeranian University of Technology, Szczecin, Poland

Abstract. The paper studies the problem of controlling the orientation of an Earth observation satellite with a single solar panel. For a simplified model of a satellite we develop a control algorithm which allows to asymptotically track a harmonically changing reference signal. As a mathematical tool we employ the general regulation theory. The obtained theoretical results are successfully tested on a numerical example.

Keywords: Satellite attitude control · Asymptotic tracking · General regulation · Error feedback controller

1 Introduction

In the era of Space 4.0 the satellite technology has become wider accessible. The current paper concentrates on applying the general regulation theory [3] to control an orientation of one of the axes of the satellite with a solar panel.

1.1 The Plant Equations of Motion

In theory the satellite with a panel show an infinite number of oscillation modes but in the real life we can always select just a few modes which are excited by a specific operation. Usually, an observing satellite, which images a sequence of adjacent pieces of field performs a periodical motion around a constant axis. In such a case, only one mode is dominant. In this case the satellite can be interpreted as two rigid bodies interconnected with a viscoelastic element and rotating around a constant axis, as shown in Fig. 1.

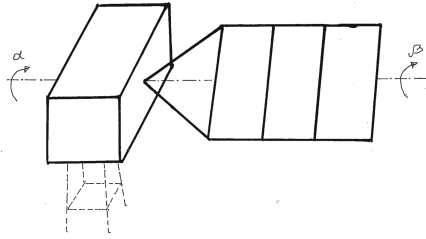


Fig. 1. Satellite with a solar panel

On the base of the above motivation we assume that the satellite with a panel is described by the following system of two second order differential equations

$$\begin{cases} I\ddot{\alpha}(t) = k(\beta(t) - \alpha(t)) + b(\dot{\beta}(t) - \dot{\alpha}(t)) + \tau(t), \\ p\ddot{\beta}(t) = -k(\beta(t) - \alpha(t)) - b(\dot{\beta}(t) - \dot{\alpha}(t)), \end{cases} \quad (1)$$

where $\tau(t)$ is a driving torque (input or control), $\alpha(t)$ is a satellite angular displacement (output), $\beta(t)$ is a panel angular displacement (unmeasured signal), I is a satellite rotational inertia, p is a panel rotational inertia, k is a stiffness coefficient, b is a friction coefficient. According to the physical meaning of the model parameters we assume that $I > 0, p > 0, k > 0$ and $b \geq 0$. In this paper we develop a control algorithm which makes the angular displacement α to follow a prescribed periodically changing reference signal α_r . Using the general regulation theory in the state space setup we derive the so-called error feedback controller.

1.2 Preliminary Formulation of the Control Problem

We assume we are given the *reference signal*

$$\alpha_r(t) = a \sin \omega t, \quad a > 0, \quad \omega > 0, \quad (2)$$

and define the *control error* $e(t) = \alpha(t) - \alpha_r(t)$. The *control goal* is to asymptotically track the reference signal $\alpha_r(t)$ by the plant output $\alpha(t)$, i.e.

$$\lim_{t \rightarrow \infty} e(t) = 0, \quad (3)$$

and this goal is to be achieved by the following *dynamic error feedback controller*

$$\Sigma_K : \begin{bmatrix} \dot{x}_K(t) \\ \tau(t) \end{bmatrix} = \begin{bmatrix} A_K & B_K \\ C_K & 0 \end{bmatrix} \begin{bmatrix} x_K(t) \\ e(t) \end{bmatrix}, \quad (4)$$

where $(x_K(t))_{t \geq 0} \subset \mathbb{R}^{n_K}$, n_K is the order of the controller and the error $e(t)$ is a *measured output* - the only signal available to the controller. Since we mainly

employ the state space methods we start by introducing the natural *state variables* $x_1 = \alpha$, $x_2 = \beta$, $x_3 = \dot{\alpha}$, $x_4 = \dot{\beta}$ and get the *plant state space model*

$$\Sigma_G : \begin{cases} \dot{x}_1(t) = x_3(t) \\ \dot{x}_2(t) = x_4(t) \\ \dot{x}_3(t) = -\frac{k}{I}x_1(t) + \frac{k}{I}x_2(t) - \frac{b}{I}x_3(t) + \frac{b}{I}x_4(t) + \frac{1}{I}\tau(t) \\ \dot{x}_4(t) = \frac{k}{p}x_1(t) - \frac{k}{p}x_2(t) + \frac{b}{p}x_3(t) - \frac{b}{p}x_4(t) \\ \alpha(t) = x_1(t), \end{cases} \quad (5)$$

which can be written as

$$\Sigma_G : \begin{bmatrix} \dot{x} \\ \alpha \end{bmatrix} = \begin{bmatrix} A & B \\ C & 0 \end{bmatrix} \begin{bmatrix} x \\ \tau \end{bmatrix}, \quad (6)$$

where $(x(t))_{t \geq 0} \subset \mathbb{R}^4$ and the *state space matrix* is given by

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} A & B \\ C & 0 \end{bmatrix} = \left[\begin{array}{cccc|c} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ -\frac{k}{I} & \frac{k}{I} & -\frac{b}{I} & \frac{b}{I} & \frac{1}{I} \\ \frac{k}{p} & -\frac{k}{p} & \frac{b}{p} & -\frac{b}{p} & 0 \\ \hline 1 & 0 & 0 & 0 & 0 \end{array} \right]. \quad (7)$$

For the plant Σ_G the *controllability* of (A, B) can be checked by using the *controllability matrix* W and the *observability* of (C, A) - by using the *observability matrix* V . Namely

$$\det W = -\frac{k^2}{I^4 p^2} \neq 0, \quad \det V = -\frac{k^2}{I^2} \neq 0, \quad (8)$$

for all values of parameters I , p , k and b with physical meaning. Actually, the control system we design is the *unit feedback control system*, shown in Fig. 2.

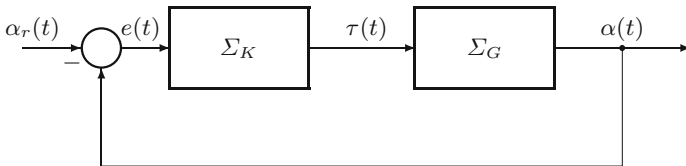


Fig. 2. Unit feedback control system

2 Precise Statement and Solution of the Control Problem

The essential feature of the approach based on the general regulation theory [1, 3] is that we assume the *reference signal* is generated by a known dynamical system, called the *exosystem*, which is aggregated with the plant model.

2.1 Reference Signal

We assume the *reference signal* of the form $\alpha_r(t) = a \sin(\omega t + \varphi)$, so it can be generated by Σ_r of the form

$$\Sigma_r : \begin{cases} \dot{r}_1(t) = r_2(t), & r_1(0) = a \sin \varphi, \\ \dot{r}_2(t) = -\omega^2 r_1(t), & r_2(0) = a\omega \cos \varphi, \\ \alpha_r(t) = r_1(t), \end{cases} \quad (9)$$

where $\omega > 0$ is known and $a \in \mathbb{R}$ and $\varphi \in \mathbb{R}$ may be unknown. By introducing

$$r(t) := \begin{bmatrix} r_1(t) \\ r_2(t) \end{bmatrix}, \quad (r(t))_{t \geq 0} \subset \mathbb{R}^2, \quad (10)$$

we get

$$\Sigma_r : \begin{cases} \dot{r}(t) = Sr(t), & r(0) = r_0, \\ \alpha_r(t) = Tr(t), \end{cases} \quad (11)$$

with the *state space matrix*

$$\begin{bmatrix} S \\ T \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -\omega^2 & 0 \\ 1 & 0 \end{bmatrix}, \quad (12)$$

with eigenvalues $\sigma(S) = \{j\omega\} \cup \{-j\omega\} \subset j\mathbb{R}$. The pair (T, S) is *observable* since the observability matrix V_r of Σ_r satisfies

$$\det V_r = \det \begin{bmatrix} T \\ TS \end{bmatrix} = 1 \neq 0. \quad (13)$$

2.2 The Extended Plant Σ_e

If we put together (6), (10) we obtain an *extended plant* Σ_e

$$\Sigma_e : \begin{cases} \dot{x}(t) = Ax(t) + B\tau(t), & x(0) = x_0, \\ \dot{r}(t) = Sr(t), & r(0) = r_0, \\ e(t) = Cx(t) - Tr(t), \end{cases} \quad (14)$$

and if we introduce an *extended state* $(\begin{bmatrix} x(t) \\ r(t) \end{bmatrix})_{t \geq 0} \subset \mathbb{R}^6$, then

$$\Sigma_e : \begin{bmatrix} \dot{x} \\ \dot{r} \\ e \end{bmatrix} = \begin{bmatrix} A_e & B_e \\ C_e & 0 \end{bmatrix} \begin{bmatrix} x \\ r \\ \tau \end{bmatrix}, \quad (15)$$

with the state space matrix

$$\begin{bmatrix} A_e & B_e \\ C_e & 0 \end{bmatrix} = \begin{bmatrix} A & 0 & B \\ 0 & S & 0 \\ C & -T & 0 \end{bmatrix}. \quad (16)$$

For $b > 0$ the plant Σ_e , or the pair (C_e, A_e) , is observable since

$$\det V_e = \det \begin{bmatrix} C_e \\ C_e A_e \\ \vdots \\ C_e A_e^5 \end{bmatrix} = -\frac{k^2 \omega^2}{I^2} \left(\omega^2 \left(\frac{b}{p} + \frac{b}{I} \right)^2 + \left(\frac{k}{p} + \frac{k}{I} - \omega^2 \right)^2 \right) \neq 0. \quad (17)$$

If $b = 0$, then Σ_e is observable if and only if $\frac{k}{p} + \frac{k}{I} \neq \omega^2$. Interconnecting (14) and (4) we obtain the *closed loop system* Σ_{cl}

$$\Sigma_{cl} : \begin{cases} \dot{x} = Ax + BC_K x_K, & x(0) = x_0, \\ \dot{x}_K = A_K x_K + B_K Cx - B_K Tr, & x_K(0) = x_{K0}, \\ \dot{r} = Sr, & r(0) = r_0, \\ e = Cx - Tr, \end{cases} \quad (18)$$

i.e.

$$\Sigma_{cl} : \begin{bmatrix} \dot{x} \\ \dot{x}_K \\ \dot{r} \\ e \end{bmatrix} = \begin{bmatrix} A & BC_K & 0 \\ B_K C & A_K & -B_K T \\ 0 & 0 & S \\ C & 0 & -T \end{bmatrix} \begin{bmatrix} x \\ x_K \\ r \end{bmatrix}, \quad \begin{bmatrix} x(0) \\ x_K(0) \\ r(0) \end{bmatrix} = \begin{bmatrix} x_0 \\ x_{K0} \\ r_0 \end{bmatrix}. \quad (19)$$

We require the controller (4) to guarantee that the closed loop system Σ_{cl} satisfies the following two conditions:

(IS) : *Internal stability*, which means that for $r(0) = r_0 = 0$ ($r(t) \equiv 0$) and all $x(0) = x_0, x_K(0) = x_{K0}$ we have

$$\lim_{t \rightarrow \infty} \begin{bmatrix} x(t) \\ x_K(t) \end{bmatrix} = 0.$$

(AT) : *Asymptotic tracking* (called *regulation*), which means that for all $r(0) = r_0$ and all $x(0) = x_0, x_K(0) = x_{K0}$ we have $\lim_{t \rightarrow \infty} e(t) = 0$.

It follows from (19) that (IS) holds if and only if

$$\sigma \left(\begin{bmatrix} A & BC_K \\ B_K C & A_K \end{bmatrix} \right) \subset \mathbb{C}_-. \quad (20)$$

In turn, (AT) requires a deeper study. Before we proceed with the general error feedback controller it is convenient to start with the simple state feedback case and this is done in the next subsection.

2.3 Static State Feedback Controller

We first consider the *static state feedback controller* for the plant Σ_e (see (14))

$$\tau = -Fx + Qr = -[f_1 \ f_2 \ f_3 \ f_4] \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} + [q_1 \ q_2] \begin{bmatrix} r_1 \\ r_2 \end{bmatrix}. \quad (21)$$

The substitution of (21) into (14) leads to the *closed loop system* Σ_{cl}

$$\Sigma_{cl} : \begin{cases} \dot{x} = (A - BF)x + BQr, & x(0) = x_0, \\ \dot{r} = Sr, & r(0) = r_0, \\ e = Cx - Tr, \end{cases} \tag{22}$$

i.e.

$$\Sigma_{cl} : \begin{bmatrix} \dot{x} \\ \dot{r} \\ e \end{bmatrix} = \begin{bmatrix} A - BF & BQ \\ 0 & S \\ C & -T \end{bmatrix} \begin{bmatrix} x \\ r \end{bmatrix}, \quad \begin{bmatrix} x(0) \\ r(0) \end{bmatrix} = \begin{bmatrix} x_0 \\ r_0 \end{bmatrix}. \tag{23}$$

It is clear that in this case (IS) of Σ_{cl} is equivalent to the fact $\sigma(A - BF) \subset \mathbb{C}_-$. Since the pair (A, B) is *controllable* (see (8)) we can freely assign eigenvalues of $A - BF$. The main result of the paper is as follows.

Theorem 1. *If $b > 0$ or $b = 0$ and $\frac{k}{p} \neq \omega^2$, then there exists a static state feedback controller (21) such that for Σ_{cl} the internal stability (IS) and asymptotic tracking (AT) hold. One such controller is given by the formula*

$$\tau = [-F \ \Gamma + F\Pi] \begin{bmatrix} x \\ r \end{bmatrix}, \tag{24}$$

where $F \in \mathbb{R}^{1 \times 4}$ is such that $\sigma(A - BF) \subset \mathbb{C}_-$ and the pair (Π, Γ) , where $\Pi \in \mathbb{R}^{4 \times 2}$ and $\Gamma \in \mathbb{R}^{1 \times 2}$, is a unique solution of the regulator equation

$$\begin{cases} A\Pi - \Pi S + B\Gamma = 0, \\ C\Pi - T = 0, \end{cases} \tag{25}$$

with explicit expressions

$$\Pi = \begin{bmatrix} \pi_{11} & \pi_{12} \\ \pi_{21} & \pi_{22} \\ \pi_{31} & \pi_{32} \\ \pi_{41} & \pi_{42} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \frac{k^2 + (b^2 - pk)\omega^2}{b^2\omega^2 + (k - p\omega^2)^2} & \frac{-bp\omega^2}{b^2\omega^2 + (k - p\omega^2)^2} \\ 0 & 1 \\ \frac{bp\omega^4}{b^2\omega^2 + (k - p\omega^2)^2} & \frac{k^2 + (b^2 - pk)\omega^2}{b^2\omega^2 + (k - p\omega^2)^2} \end{bmatrix} \tag{26}$$

and

$$\begin{aligned} \Gamma &= [\gamma_1 \ \gamma_2] \\ &= \left[-\frac{\omega^2(b^2\omega^2(p+I) + kp(k-p\omega^2) + I(k-p\omega^2)^2)}{b^2\omega^2 + (k-p\omega^2)^2} \quad \frac{bp^2\omega^4}{b^2\omega^2 + (k-p\omega^2)^2} \right]. \end{aligned} \tag{27}$$

Proof. The results of the theorem with the exception for the formulas (26) and (27) can be derived from the general regulation theory [1,3]. However, the usual difficulty lies in showing *if and when* the regulator equation (25) admits a solution. As our main specific result we prove that for the extended plant Σ_e with $b > 0$ or $b = 0$ and $\frac{k}{p} \neq \omega^2$, the regulator equation (25) has a solution and this solution is unique. Moreover, we find this solution explicitly. By introducing

$$\Pi = \begin{bmatrix} \pi_{11} & \pi_{12} \\ \pi_{21} & \pi_{22} \\ \pi_{31} & \pi_{32} \\ \pi_{41} & \pi_{42} \end{bmatrix}, \quad \Gamma = [\gamma_1 \ \gamma_2], \tag{28}$$

we can rewrite (25) explicitly as the system of algebraic equations

$$\begin{bmatrix}
 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & \omega^2 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
 1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & \omega^2 & 0 & 0 & 1 & 0 & 0 & 0 \\
 0 & 0 & 1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\
 -k & 0 & k & 0 & -b & I\omega^2 & b & 0 & 1 & 0 \\
 0 & -k & 0 & k & -1 & -b & 0 & b & 0 & 1 \\
 k & 0 & -k & 0 & b & 0 & -b & p\omega^2 & 0 & 0 \\
 0 & k & 0 & -k & 0 & b & -p & -b & 0 & 0
 \end{bmatrix}
 \begin{bmatrix}
 \pi_{11} \\
 \pi_{12} \\
 \pi_{21} \\
 \pi_{22} \\
 \pi_{31} \\
 \pi_{32} \\
 \pi_{41} \\
 \pi_{42} \\
 \gamma_1 \\
 \gamma_2
 \end{bmatrix}
 =
 \begin{bmatrix}
 1 \\
 0 \\
 0 \\
 0 \\
 0 \\
 0 \\
 0 \\
 0 \\
 0 \\
 0
 \end{bmatrix}, \tag{29}$$

i.e.

$$M \text{vec}(\Pi, \Gamma) = N. \tag{30}$$

If $I > 0$, $p > 0$, $k > 0$ and $b > 0$, then after some tedious calculations we obtain $\det M = b^2\omega^2 + (k - p\omega^2)^2 \neq 0$, and if $b = 0$, then for $\frac{k}{p} \neq \omega^2$ we still get $\det M \neq 0$. In both cases we can invert M and solve the system (29) to get a unique solution $\text{vec}(\Pi, \Gamma) = M^{-1}N$, where Π and Γ are explicitly given by (26) and (27), respectively. \square

The above result is essential since it provides a basis for the development of the error feedback controller in the form (4).

2.4 Error Feedback Controller Based on a Full Order Observer

By (8) we know that for $b \geq 0$ the pair (A, B) is controllable and (C, A) is observable. In this case there always exists a controller (A_K, B_K, C_K) satisfying (20). For a controller, satisfying (IS), the general regulation theory [3] provides necessary and sufficient conditions to be satisfied for the asymptotic tracking condition (AT) to hold. Below we provide appropriate results tailored to our purposes.

Theorem 2. *If for a given controller (A_K, B_K, C_K) the condition (IS) holds, then the closed loop system Σ_{cl} satisfies the asymptotic tracking condition (AT) if and only if there exist $\Pi \in \mathbb{R}^{4 \times 2}$, $\Gamma \in \mathbb{R}^{1 \times 2}$ and $\Sigma \in \mathbb{R}^{n_K \times 2}$ such that*

$$(\text{RE}) : \begin{cases} A\Pi - \Pi S + B\Gamma = 0, \\ C\Pi - T = 0, \end{cases} \tag{31}$$

and

$$(\text{IMP}) : \begin{cases} \Gamma = C_K \Sigma, \\ \Sigma S = A_K \Sigma. \end{cases} \tag{32}$$

The relation (31) is the *regulator equation* from Theorem 1, so we denote it (RE), and the relation (32) is called the *internal model principle* [2] and hence we denote it (IMP). The latter relation reflects the fact that the dynamics of

the exosystem appears in the controller. One can use the relations (20), (31) and (32) as inspiration for developing a controller (A_K, B_K, C_K) . Instead, we can turn to the result of Theorem 1 for the static state feedback controller. However, the components of the extended state $\begin{bmatrix} x \\ r \end{bmatrix}$ cannot be measured since we only measure the error signal e , i.e. the output of Σ_e . In this case can construct a classic *full order asymptotic observer* to estimate the extended state and use the approximate state $\begin{bmatrix} \tilde{x} \\ \tilde{r} \end{bmatrix}$ in the control law

$$\tau = [-F \Gamma + F\Pi] \begin{bmatrix} \tilde{x} \\ \tilde{r} \end{bmatrix}. \tag{33}$$

We have shown (see (17)) that if $b > 0$ or $b = 0$ and $\frac{k}{p} + \frac{k}{T} \neq \omega^2$, then the plant Σ_e is observable. The classic Luenberger *asymptotic full order state observer* for Σ_e leads to a *controller* Σ_K , of the order $n_K = 6$, in the form (4)), where $x_K = \begin{bmatrix} \tilde{x} \\ \tilde{r} \end{bmatrix}$ and

$$\begin{aligned} A_K &= \begin{bmatrix} A & 0 \\ 0 & S \end{bmatrix} - \begin{bmatrix} L_1 \\ L_2 \end{bmatrix} [C \ -T] + \begin{bmatrix} B \\ 0 \end{bmatrix} [-F \Gamma + F\Pi], & B_K &= \begin{bmatrix} L_1 \\ L_2 \end{bmatrix}, \\ C_K &= [-F \Gamma + F\Pi], & D_K &= 0. \end{aligned} \tag{34}$$

One can show that for the controller (34) the conditions (IS), (RE) and (IMP) hold. Due to lack of space we omit details.

3 Numerical Simulations

In order to verify the performance of the control system of Fig.2 a series of numerical simulations have been performed using the Matlab/Simulink package. One example is presented below. We assume the following data

$$I = 1, \quad p = 0.1, \quad k = 0.15, \quad b = 0.01, \quad a = 1, \quad \omega = 2, \tag{35}$$

and get the state space matrices of Σ_G and Σ_r

$$\left[\begin{array}{c|c} A & B \\ \hline C & 0 \end{array} \right] = \left[\begin{array}{cccc|c} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ -0.15 & 0.15 & -0.01 & 0.01 & 1 \\ 1.5 & -1.5 & 0.1 & -0.1 & 0 \\ \hline 1 & 0 & 0 & 0 & 0 \end{array} \right], \quad \left[\begin{array}{c} S \\ T \end{array} \right] = \left[\begin{array}{cc} 0 & 1 \\ -4 & 0 \\ 1 & 0 \end{array} \right], \tag{36}$$

and $\alpha_r(t) = \sin 2t$. Moreover, the solution of (RE) is given by

$$\Pi = \begin{bmatrix} 1.0000 & 0 \\ -0.5898 & -0.0636 \\ 0 & 1.0000 \\ 0.2544 & -0.5898 \end{bmatrix}, \quad \Gamma = [-3.7641 \ 0.0254]. \tag{37}$$

The state feedback gain matrix $F \in \mathbb{R}^{1 \times 4}$ is chosen to minimize the functional

$$J(\tau) = w_0 \int_0^\infty x^T(t)x(t)dt + u_0 \int_0^\infty \tau^2(t)dt, \tag{38}$$

with weights $w_0 = 10$ and $u_0 = 1$, for the system $\dot{x}(t) = Ax(t) + B\tau(t)$ with $\tau(t) = -Fx(t)$. We used the MATLAB *lqr* procedure and get

$$F = [7.1263 \ -2.6542 \ 4.9695 \ 2.7116], \quad \sigma(A - BF) = \begin{bmatrix} -2.9642 \\ -0.5328 + j1.3682 \\ -0.5328 - j1.3682 \\ -1.0498 \end{bmatrix}.$$

The output error gain matrix $\begin{bmatrix} L_1 \\ L_2 \end{bmatrix} \in \mathbb{R}^{6 \times 1}$ is chosen such that $\begin{bmatrix} L_1 \\ L_2 \end{bmatrix}^T \in \mathbb{R}^{1 \times 6}$ minimizes the quadratic cost functional

$$J(\vartheta) = w_1 \int_0^\infty \xi^T(t)\xi(t)dt + u_1 \int_0^\infty \vartheta^2(t)dt, \tag{39}$$

with weights $w_1 = 100$ and $u_1 = 1$, for the system $\dot{\xi}(t) = \begin{bmatrix} A & 0 \\ 0 & S \end{bmatrix}^T \xi(t) + [C - T]^T \vartheta(t)$ with $\vartheta(t) = -\begin{bmatrix} L_1 \\ L_2 \end{bmatrix}^T \xi(t)$. Hence,

$$\begin{bmatrix} L_1 \\ L_2 \end{bmatrix} = \begin{bmatrix} 15.0514 \\ 4.6231 \\ 9.4472 \\ 6.0269 \\ -1.1772 \\ -22.2364 \end{bmatrix}, \quad \sigma\left(\begin{bmatrix} A & 0 \\ 0 & S \end{bmatrix} - \begin{bmatrix} L_1 \\ L_2 \end{bmatrix} [C - T]\right) = \begin{bmatrix} -13.9592 \\ -1.0056 \\ -0.6024 + j1.5505 \\ -0.6024 - j1.5505 \\ -0.0845 + j1.2431 \\ -0.0845 - j1.2431 \end{bmatrix}.$$

The obtained state space matrix of the controller is given by

$$\left[\begin{array}{c|c} \begin{matrix} A_K & B_K \\ C_K & 0 \end{matrix} & \end{array} \right] = \left[\begin{array}{cccccc|c} -15.0514 & 0 & 1.0000 & 0 & 15.0514 & 0 & 15.0514 \\ -4.6231 & 0 & 0 & 1.0000 & 4.6231 & 0 & 4.6231 \\ -16.7235 & 2.8042 & -4.9795 & -2.7016 & 15.0647 & 3.5643 & 9.4472 \\ -4.5269 & -1.5000 & 0.1000 & -0.1000 & 6.0269 & 0 & 6.0269 \\ 1.1772 & 0 & 0 & 0 & -1.1772 & 1.0000 & -1.1772 \\ 22.2364 & 0 & 0 & 0 & -26.2364 & 0 & -22.2364 \\ \hline -7.1263 & 2.6542 & -4.9695 & -2.7116 & 5.6175 & 3.5643 & 0 \end{array} \right].$$

The effectiveness of the method has been confirmed by numerical simulations and some of them are shown in Figs. 3, 4 and 5.

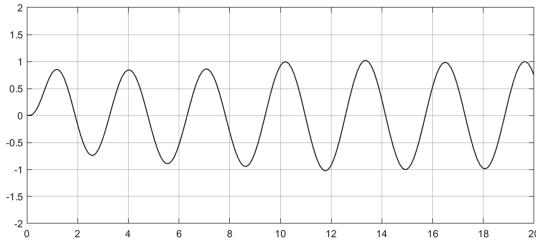


Fig. 3. The output $\alpha(t)$

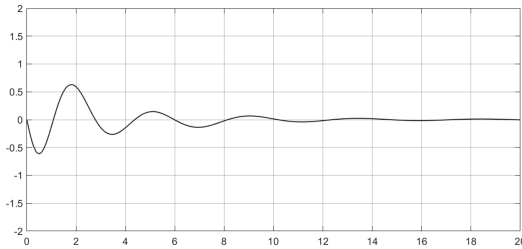


Fig. 4. The error $e(t)$

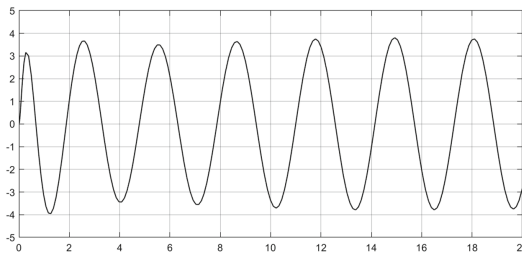


Fig. 5. The input $\tau(t)$

References

1. Francis, B.A.: The linear multivariable regulator problem. *SIAM J. Control Opt.* **15**(3), 486–505 (1977)
2. Francis, B.A., Wonham, W.M.: The internal model principle for linear multivariable regulators. *Appl. Math. Opt.* **2**(2), 170–194 (1975)
3. Saberi, A., Stoorvogel, A.A., Sannuti, P.: *Control of Linear Systems with Regulation and Input Constraints*. Springer, London (2000). <https://doi.org/10.1007/978-1-4471-0727-9>
4. Wen, T.-Y.J., Kreutz-Delgado, K.: The attitude control problem. *IEEE Trans. Autom. Control* **36**(10), 1148–1162 (1991)

Author Index

A

Abadi, Ali Soltani Sharif 329, 372
Ahsan, Muhammad 294
Alexiev, Alexander 287
Ambrożkiewicz, Bartłomiej 261
Anczarski, Jakub 261

B

Babisz, Radosław 135
Badecka, Inez 145
Barciński, Tomasz 393
Bartyś, Michał 205
Batsala, Yaroslav 269
Bauke, Dawid 173
Bieszczad, Rafal 305
Bismor, Dariusz 27, 294
Błaszczyk, Tomasz 123
Bożek, Andrzej 320
Bukowiecka, Nikola 393

C

Czyż, Zbigniew 261

D

Dereviahina, Nataliia 193
Domek, Stefan 103
Dworak, Paweł 381

E

Emirsajłow, Zbigniew 393

F

Figwer, Jarosław 49
Foit, Krzysztof 123

G

Główka, Teresa 59
Goldschmidt, Nico 237

H

Hlad, Ivan 269
Hosseinabadi, Pooyan Alinaghi 329, 372

I

Inkin, Olexander 193

J

Jabłoński, Andrzej 39
Jabłoński, Karol 27
Jachowicz, Mikołaj 261
Janusz, Wojciech 173
Jastrzębski, Marcin 349

K

Kabziński, Jacek 349
Kampa, Adrian 123
Karim Afshar, Keyvan 337
Karimi, Masoud 145
Karwatka, Jakub 135
Kasprzyk, Jerzy 113
Konieczny, Jarosław 337
Koper, Piotr 135
Korbicz, Józef 250
Korus, Lukasz 39
Korzeniowski, Roman 337
Kościelny, Jan Maciej 205
Kozyra, Andrzej 135
Král, Ladislav 227
Krauze, Piotr 173
Król, Szymon 381
Kulik, Jakub 123
Kuts, Yuriy 287

L

Liaskovska, S. Y. 3
Lipczyńska, Aleksandra 135
Loska, Jacek 113

Łukowicz, Krzysztof 123
Łydek, Paweł 145
Lysenko, Iuliia 287
Ławryńczuk, Maciej 183, 361

M

Madej, Damian 135
Martyń, Y. V. 3
Michalczyk, Małgorzata I. 216
Milanowski, Hubert 305
Mirchev, Yordan 287
Mitkowski, Wojciech 158
Molokanova, V. M. 86
Mróz, Miłosz 123
Mudka, Zbigniew 173

N

Nebeluk, Robert 361
Nowak, Julia 123
Nowakowski, Konrad 135
Nykolyn, Petro 69
Nykolyn, Uliana 69

O

Olszówka, Przemysław 173
Oprzędkiewicz, Krzysztof 158
Ordys, Andrew 329, 372

P

Płaczek, Marek 173
Pazera, Marcin 250
Pilat, Adam Krzysztof 305
Przybyła, Grzegorz 173
Pukocz, Przemysław 145
Punčochář, Ivo 227

R

Rodzik, Dariusz 123
Rosół, Maciej 158
Rydel, Marek 16

S

Sadovenko, Ivan 193
Schulte, Horst 237
Sękała, Agnieszka 123
Sikora, Bartłomiej 305
Skulimowski, Andrzej M. J. 145
Stączek, Paweł 261
Stanisławski, Rafał 16
Stetter, Ralf 277
Świder, Zbigniew 320
Szulc, Michał 113

T

Tomczok, Dominik 135
Trybus, Leszek 320
Turek, Jakub 173

U

Uchanin, Valentyn 287
Uciński, Dariusz 76

W

Witański, Marek 123
Witeczak, Marcin 250, 277
Wychowański, Michał 173
Wyciśłok, Artur 173

Z

Zarzycki, Krzysztof 183
Zawiślak, Rafał 349
Ziaja, Maciej 173
Zosgórnik, Szymon 173
Żebrowski, Patryk 123
Żmudka, Zbigniew 173